6-1999

Use of Bioinformatics in the Analysis of Chemotaxis Proteins

Sean Andrew Bulloch

LOMA LINDA UNIVERSITY

Graduate School

USE OF BIOINFORMATICS IN THE ANALYSIS OF CHEMOTAXIS PROTEINS

by

Sean Andrew Bulloch

A Thesis in Partial Fulfillment

of the Requirements for the Degree Master of

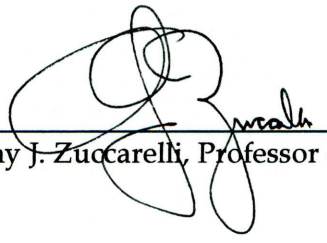
Science in Microbiology and Molecular Genetics

June 1999

Each person whose signature appears below certifies that this thesis in their opinion is adequate, in scope and quality, as a thesis for the degree Master of Science in Microbiology and Molecular Genetics.

Igor B. Zhulin, Assistant Professor of Microbiology and Molecular Genetics, Chairperson

Mark S. Johnson, Associate Research Professor of Microbiology and Molecular Genetics



Anthony J. Zuccarelli, Professor of Microbiology and Molecular Genetics

ACKNOWLEDGEMENTS

I would like to express my appreciation to the individuals who helped me complete this research. I am very grateful to Dr. Igor Zhulin for serving as my mentor and for his supervision, instruction and support that made this work possible. I would also like to thank Drs. Mark Johnson and Anthony Zuccarelli for their service as my committee members.

I am especially indebted to Dr. Robert Bourret, Dr. Jonathan Eisen and Dr. John Kirby for their collaboration and contributions to this work. I am grateful to Dr. Gladys Alexandre for her hours of instruction in teaching me a wide variety of laboratory techniques as a supplement to many hours of computer work. I also thank Dr. Judith Armitage for providing me with protein sequences before publication. I dedicate this manuscript to my parents for their continual love and support throughout my educational experience.

TABLE OF CONTENTS

CHAPTER FOUR

LIST OF FIGURES

Figures Page

LIST OF TABLES

Table Page

LIST OF ABBREVIATIONS

ATP	-	Adenosine triphosphate
BLAST	-	Basic Local Alignment Search Tool
CW	-	Clockwise rotation
CCW	-	Counter-clockwise rotation
DAS	-	Dense Alignment Surface method
FAD	-	Flavin adenine dinucleotide
HCD	-	Highly conserved domain
MCP	-	Methyl-accepting chemotaxis protein
NCBI	-	National Center for Biotechnology Information
PAS	-	Sensory domain present in *Drosophila* period clock protein (PER), vertebrate Aryl hydrocarbon receptor nuclear translocator (ARNT) and *Drosophila* single-minded protein (SIM)
PSI-BLAST	-	Position Specific Iterated - Basic Local Alignment Search Tool
SAM	-	S-adenosylmethionine
TIGR	-	The Institute for Genomic Research

ABSTRACT

USE OF BIOINFORMATICS IN THE ANALYSIS OF CHEMOTAXIS PROTEINS

by

Sean Andrew Bulloch

Bacterial chemotaxis is one of the best-known signal transduction systems. Levels of attractants and repellents are sensed in the surrounding environment by various chemoreceptors. The signal is passed to the excitation pathway via the transfer of a phosphoryl group from the receptor-associated histidine kinase CheA to the response regulator CheY. Phospho-CheY binds to the flagellar motor switching the direction of rotation of the flagella and thus allowing the cell to move towards or away from the attractant or repellent. The phosphatase CheZ removes the phosphoryl group from phospho-CheY restoring default flagellar rotation. Adaptation to stimuli occurs via addition of methyl groups to the receptor by the CheR methyltransferase and their removal by the CheB methylesterase.

In order to learn more about the diversity of signal transduction, we used bioinformatics to analyze two of the key proteins in the chemotaxis pathway: the single-domain response regulator CheY and multi-domain methyl-accepting chemotaxis proteins (MCPs). Database searches revealed more than 50 known and putative CheY proteins and CheY-like domains with alignment analysis revealing overall conservation in protein folding. From published papers, critical residues were identified that played important roles in phosphorylation and interaction with CheA, CheZ and FliM proteins.

Topological studies classified all known and putative MCPs into six distinct classes. Analysis of the C-terminal signaling domain showed that the two methylation regions (K1 and R1) as well as the highly conserved domain (HCD) had a similar fold and a high degree of conservation. PSI-BLAST analysis showed a statistically significant relationship ($E < .001$) between the N-terminal sensing domain of bacterial chemporeceptors and the sensing domains from histidine and serine/threonine kinases.

CHAPTER ONE
INTRODUCTION

A. Bacterial chemotaxis

In order to respond to changes in external and internal environments, organisms need a way of transmitting information to the appropriate location in a cell. One of the simplest methods of information transmission is a two-component signaling pathway. All two-component signaling pathways consist of a sensor histidine kinase (transmitter) and an aspartate response regulator (receiver) (Appleby et al., 1996) (Figure 1). Two-component systems in prokaryotes can regulate a variety of functions including osmotic pressure (Tanaka et al., 1998) and sporulation (Maeda et al., 1994). Similar pathways have also been found in eukaryotes, such as *Neurspora crassa*, where hyphal development is regulated by a two-component signaling system (Alex et al., 1996).

The bacterial chemotaxis pathway is one of the best-characterized two-component signaling pathways and is unique in that it regulates protein-protein interactions, not transcription (Table 1). Pfeffer first described bacterial chemical sensing in the 1880s when he observed that bacteria tended to migrate towards oxygen-producing chloroplasts and disperse when exposed to other toxic chemicals. He also noted how certain chemicals could act as both attractants and repellents (Pfeffer, 1883). Now, most of the underlying processes involved in this chemotactic process have been identified using the chemosensory pathways in *Escherichia coli* and *Salmonella typhimurium*.

Figure 1. Basic scheme of the information processing in a two-component signaling pathway. The components common to both prokaryotic and eukaryotic signaling pathways shown are: a) the sensor kinase, containing both a sensor domain and a histidine kinase transmitter that can be autophosphorylated and b) the response regulator, which receives the phosphoryl group from the transmitter. The response regulator controls physiological signals including gene transcription and enzyme/motor regulation. Some response regulators can de-phosphorylate themselves; but other pathways contain a phosphatase that can speed up the process or can act as a pathway regulator. P, phosphate group; H, histidine residue; D, aspartate residue. Adapted from (Appleby et al., 1996).

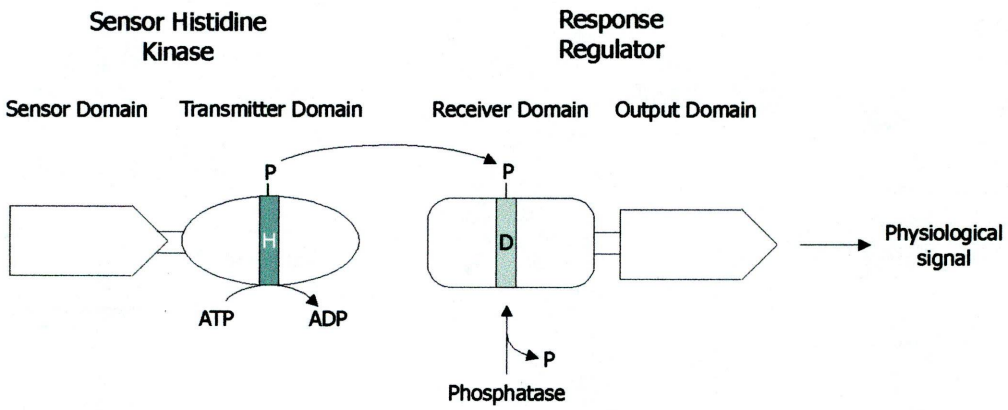

Sensor Histidine Kinase
Response Regulator
Sensor Domain
Transmitter Domain
Receiver Domain
Output Domain
P
P
H
D
Physiological signal
ATP
ADP
P
Phosphatase

Table 1. Components of the *E. coli* chemotaxis pathway

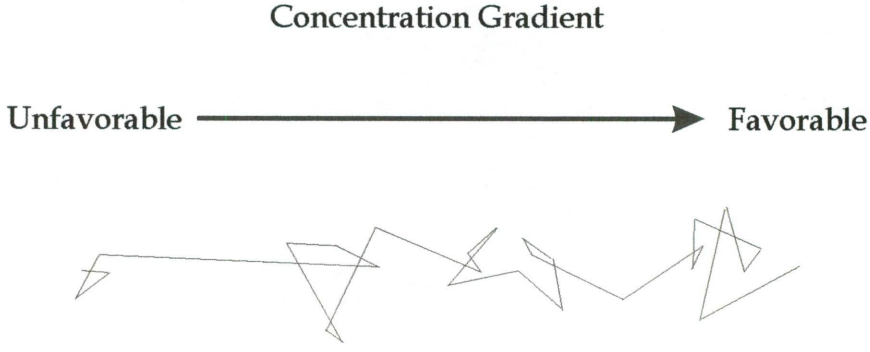
Protein	Biochemical Function
CheA	Histidine Kinase. Phosphorylation of CheY and CheB
CheB	Methylesterase. Demethylation of receptor, receives phosphoryl group from CheA
CheR	Methyltransferase. Methylation of receptor, receives methyl groups from S-adenosylmethionine
CheW	Docking protein that connects CheA with the receptors
CheY	Response regulator. Receives phosphoryl group from CheA, phospho-CheY binds to FliM (flagellar switch protein)
CheZ	Phosphatase. Removes the phosphoryl group from CheY
Methyl-accepting chemotaxis protein	Receptor (transducer). Ligand binding, signal transmission and regulation of CheA phosphorylation

Bacteria swim by rotating their flagella, powered by proton motive force (Berg and Anderson, 1973; DeRosier, 1998). *E. coli* contains six to eight left-handed helical flagella. As a result, when the flagella are rotating in a counter-clockwise direction (CCW), all of the force is pointed inward causing the flagella to form a bundle. This bundle formation allows the cell to swim in a smooth manner. If the flagella are rotated in a clockwise (CW) manner the force is point outward. Thus the flagella will fly apart causing the cell to tumble (Berg, 1993).

Normally, in the absence of any chemical gradient, cells exhibit a random 3-dimensional walk. The cell will swim in a smooth straight line during CCW flagellar rotation then tumble during CW rotation. Tumbling causes the cell to reorient and swim off in a new random direction. In the presence of a chemical gradient, control of flagellar rotation is influenced via the chemotaxis sensory pathway. When the cell is moving towards favorable chemical concentrations the frequency of tumbling is decreased while the duration of smooth swimming is increased. Alternately, when the cell is moving away from favorable conditions, the tumbling frequency is increased as to increase the probability of movement towards favorable conditions again. This influence introduces a bias into the cell's random walk towards favorable conditions (Berg and Brown, 1972) (Figure 2).

Cells sense different chemical stimuli in the environment via various transmembrane receptors. In *E. coli* there are four transmembrane chemotaxis receptors: Trg (ribose and galactose), Tsr (serine), Tar (aspartate) and Tap (dipeptides) (Boyd et al., 1981; Stock and Surette, 1996; Grebe and Stock, 1998). A fifth sensor, Aer, has recently

Figure 2. Biased random walk of a bacterium. Normal bacteria movement is a random process. In the presence of a concentration gradient the random movement becomes biased. Exposure of the bacterium to a favorable condition (i.e. increased concentration of attractants or decrease in repellent), suppresses the tumbling frequency. The bacterium swims in a smooth fashion longer towards the more favorable conditions. If the cell moves towards an unfavorable condition (i.e. increased concentration of a repellent or decrease in attractant), the tumbling is enhanced. This enhancement allows for a greater chance of movement towards favorable conditions again. Adapted from (Berg, 1993).

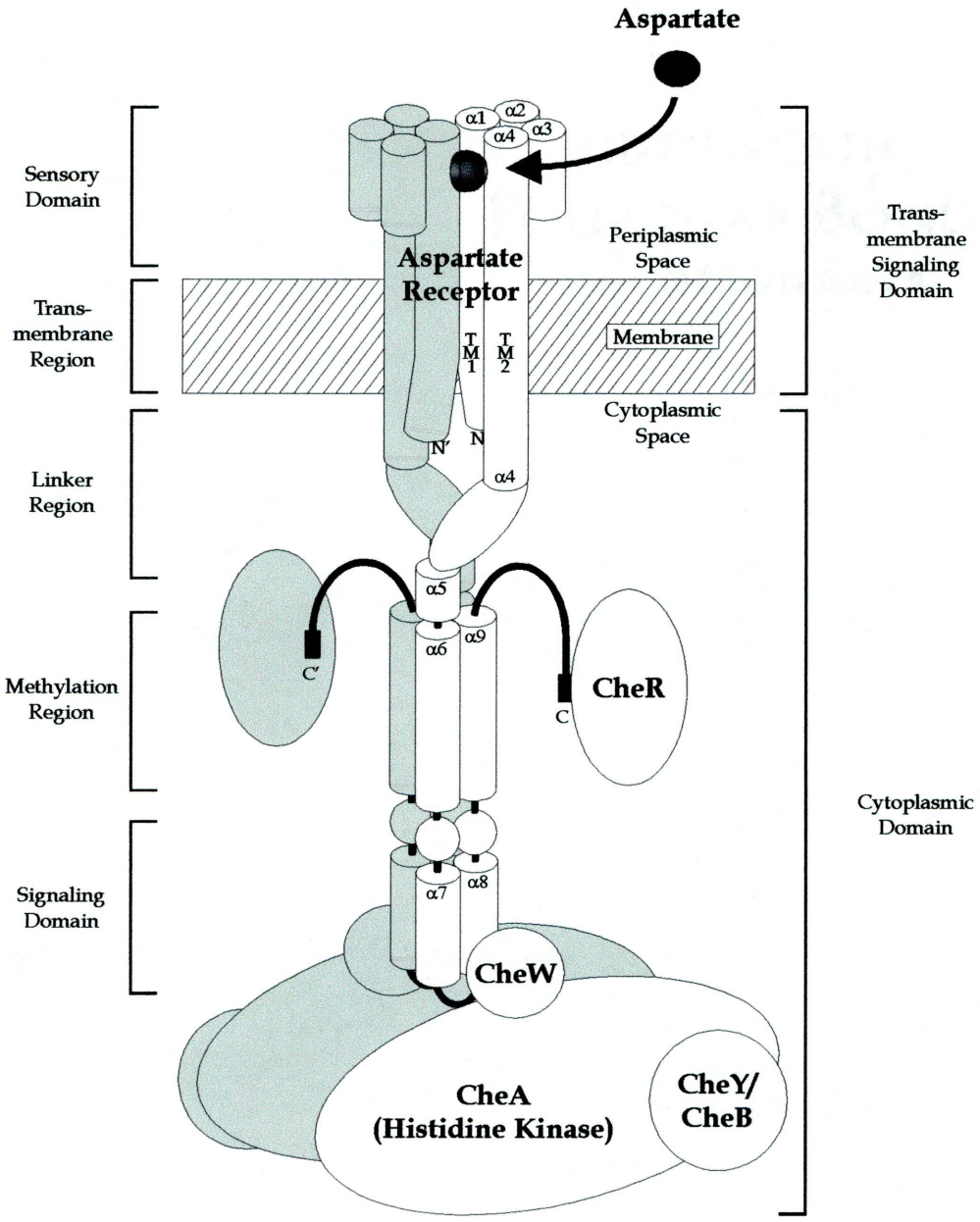

Concentration Gradient
Unfavorable
Favorable

been identified and shown to be involved in both aerotaxis and chemotaxis regulation. Aer is unique in the *E. coli* chemotaxis receptor family in that it lacks a periplasmic sensing domain. Instead, its sensing domain is located in the cytoplasm and contains a PAS domain that associates with FAD (Rebbapragada et al., 1997; Bibikov et al., 1997). PAS domains act as sensors for light, redox potential and oxygen in various organisms (Taylor and Zhulin, 1999).

The four chemotactic receptors each contain two transmembrane segments, a periplasmic sensing domain and a cytoplasmic signaling domain (Yeh et al., 1993) (Figure 3). In order to receive and transmit a signal across the membrane, the receptors dimerize to form active stable complexes (Milligan and Koshland, 1988). Crystal structure analysis of Tar revealed that the sensing domain is comprised of four α-helix bundles with two symmetrical aspartate-binding sites located at the extreme end of the domain (Milburn et al., 1991; Stoddard and Koshland, 1992).

The aspartate molecule preferentially binds in an asymmetrical fashion to one of the two binding pockets causing a displacement in the α3 and α4 helices (Yeh et al., 1996; Milburn et al., 1991). Based on crystallography studies, the α3 helix was shown not to span the membrane leaving the second transmembrane α4 helix (α4/TM2) as the only option for signal propagation (Milburn et al., 1991). Artificially engineered disulfide bonds determined that the first transmembrane α-helix transmembrane (α1/TM1) does not play a role is transmembrane signaling. It is the α4/TM2 which undergoes a piston like motion, inducing a conformational change in the signaling domain (Chervitz et al., 1995; Chervitz and Falke, 1996). A slight tilt in the signaling domain was observed with

Figure 3. Diagram of the aspartate receptor (Tar) complex from *E. coli*. Upon aspartate binding to the sensory domain, the signal is transmitted down the α4/TM2 to the signaling domain. Some of the chemotaxis pathway components are associated with the receptor including the docking protein CheW and the histidine kinase CheA. Other components are usually soluble and only transiently associated with the receptor including the methyltransferase CheR, the methylesterase CheB, and the response regulator CheY. Adapted from (Falke et al., 1997).


Aspartate
Sensory Domain
Trans-membrane Region
Linker Region
Methylation Region
Signaling Domain
Trans-membrane Signaling Domain
Cytoplasmic Domain
α1
α2
α3
α4
Aspartate Receptor
Periplasmic Space
Membrane
TM1
TM2
Cytoplasmic Space
N'
N
α4
α5
α6
α9
C'
C
CheR
α7
α8
CheW
CheA (Histidine Kinase)
CheY/ CheB

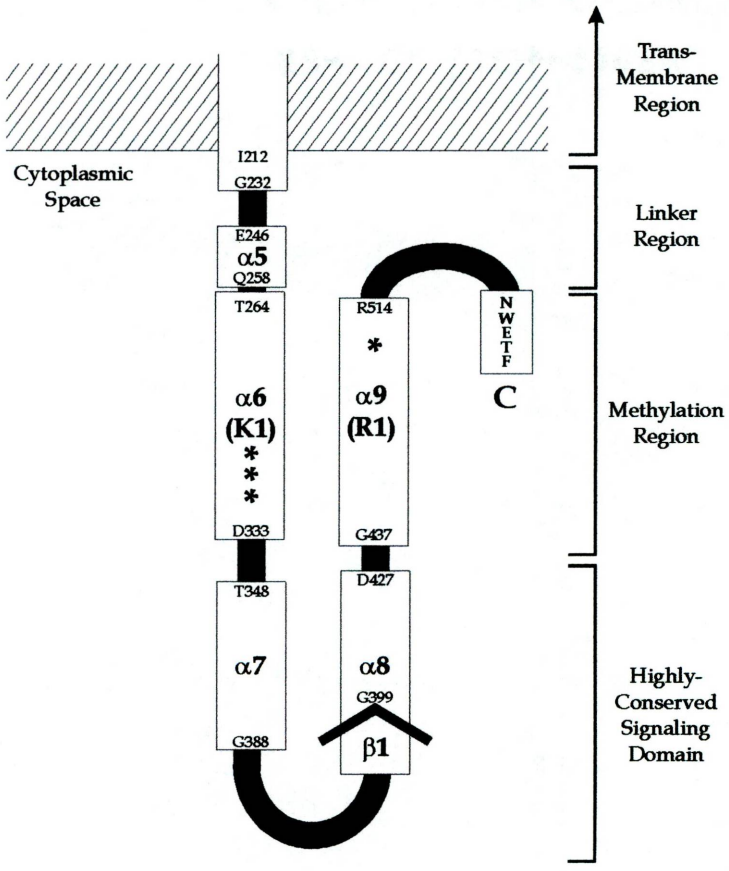
a ligand bound when compared with the apo (empty) receptor. Due to the relative stiffness of the helix, a piston like motion would transmit a signal more efficiently than a tilt. However, a 5° tilt could displace the cytoplasmic portion of the signaling helix up to 6 Å, yielding a greater signal transmission (Chervitz and Falke, 1996).

The signaling domain is located in the cytoplasmic space and comprises the entire C-terminal of the protein. Current structural models suggest that this domain is comprised of ten α-helices (5 per homeodimer) and 2 β-sheets (1 per homeodimer) which can be classified into three distinct functional groups: the linker region, the methylation regions (termed K1 and R1) and the highly conserved domain (HCD) (Stock et al., 1991; Le Moual and Koshland, 1996; Danielson et al., 1997) (Figure 4).

The linker region is a small domain approximately 30 residues in length. This region links the α4/TM2 to the signaling domain (Le Moual and Koshland, 1996). Lock-on and -off mutations interrupt signaling suggesting a role in signal transmission (Ames et al., 1988). Additionally dimers that lack one of the cytoplasmic signaling units yet retain the linker region can still transmit a signal (Tatsuno et al., 1996; Gardina and Manson, 1996). This along with recent studies revealing a coiled coil motif in the region provides further evidence for a possible role in signaling (Singh et al., 1998).

Following the linker region are two methylation domains (Terwilliger et al., 1983). The first domain (the K1 region) lies on the α6 helix while the second (the R1 region) lies in an antiparallel direction on the α9 helix (Kehry and Dahlquist, 1982). Both of these methylation domains play a critical role in the adaptation branch of the chemotaxis pathway. The adaptation branch allows a cell to recognize background

Figure 4. Predicted secondary structure of the cytoplasmic domain of the Tar receptor in *E. coli*. The cytoplasmic domain is made up of five α-helices (α5 - α9) and one β-sheet (β1). Adaptation of the receptor takes place via the interaction of CheR/CheB and the K1/R1 methylation regions. The CheR binding site (NWETF) is located at the extreme C-terminal end of the receptor. CheW and CheA bind to the receptor via the highly conserved signaling domain. Asterisks (*) represent the individual methylation sites (K1: Q295, E302, Q309; R1: Q491), and curved or straight lines represent random coils or turns. Adapted from (Le Moual and Koshland, 1996).

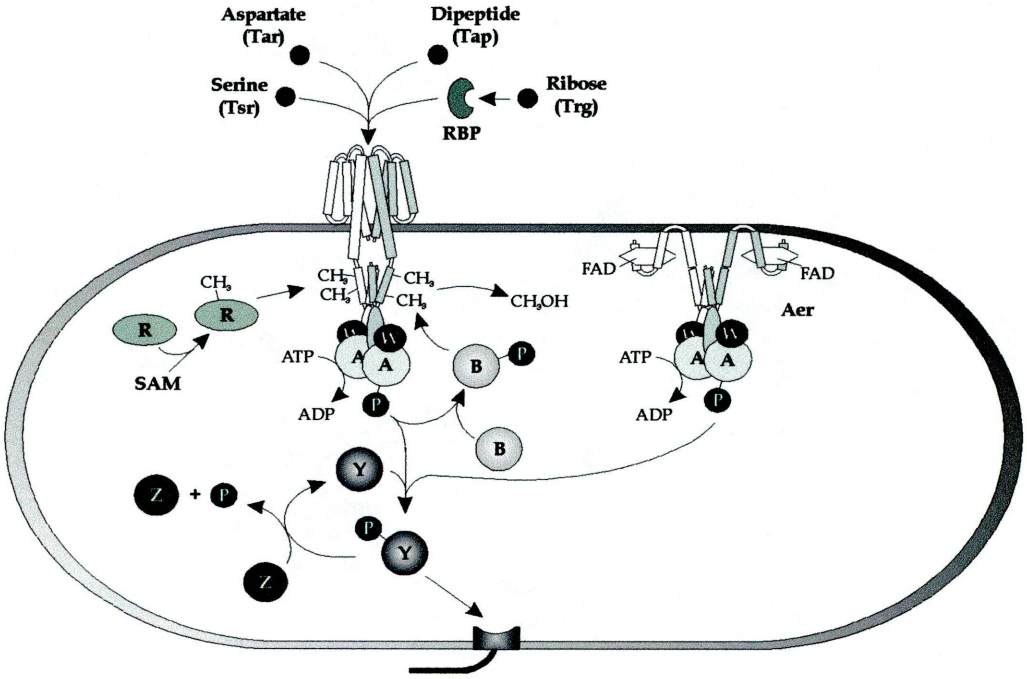

Trans-
Membrane
Region
Linker
Region
Methylation
Region
Highly-
Conserved
Signaling
Domain
Cytoplasmic
Space
I212
G232
E246
α5
Q258
T264
α6
(K1)
D333
T348
α7
G388
R514
α9
(R1)
G437
D427
α8
G399
β1
N
W
E
T
F
C

stimuli increases so that it may continue to exhibit chemotaxis behavior in a gradient regardless of the background concentration (Stock and Surette, 1996).

The specific methylation sites are glutamate side chains located in the following consensus: Glu-**Glx**-X-X-Ala-Ser/Thr where X represents any amino acid, Glx represents either a glutamate or glutamine residue and the bold residue is the methylated site (Terwilliger et al., 1986). The number of individual methylation sites in the K1/R1 regions varies from one receptor to another (Kehry and Dahlquist, 1982; Le Moual and Koshland, 1996; Danielson et al., 1997). In Tar, there are three methylation sites in the K1 region and one methylation site in the R1 region (Nowlin et al., 1987).

The glutamate residues are methylated by the methyltransferase CheR (Kort et al., 1975; Springer and Koshland, 1977) with methyl groups obtained from S-adenosylmethionine (SAM) (Terwilliger et al., 1986; Simms et al., 1987). CheR binds to a highly conserved C-terminal motif (Asn-Trp-Glu-Thr/Ser-Phe) corresponding to the last five residues present in Tsr and Tar but absent in Trg and Tap (Wu et al., 1996). Current theory suggests that a bound CheR is shared with those receptors that lack the binding motif (Okumura et al., 1998). Upon methylation, the activity of the histidine kinase CheA is increased. CheA is autophosphorylated by ATP (Hess et al., 1988) and the phosphoryl group is transferred to the methylesterase CheB (Djordjevic et al., 1998) (Li et al., 1995). The phospho-CheB (CheB~P) removes the methyl groups bound to the receptors (Stock and Surette, 1996; Stock and Koshland, 1978) (Figure 5). The rate of methylation versus the rate of demethylation is defined as a cell's steady state (Stock and Surette, 1996). Since this process of reversible methylation has been identified

Figure 5. Receptor mediated chemotaxis pathway in *E. coli*. The ligand binds to the receptor, transmitting the signal through the transmembrane region to the cytoplasmic signaling domain. The histidine kinase CheA, attached to the receptor by the CheW docking protein, transfers a phosphoryl group to two response regulators, CheY and CheB. Aer is similar to the other receptors in that it binds a CheA/CheW complex; but it does not contain a periplasmic ligand-binding domain. Instead, Aer obtains its signal through interactions with the electron transport system by the FAD-containing PAS-domain. CheR is responsible for the addition of methyl groups to the receptor. CheB removes methyl groups from the receptor when phosphorylated. Together they guide adaptation to external signals. Aer is not methylated. The pathways converge upon the phosphorylation of CheY. Phospho-CheY binds to the flagella motor, altering both flagellar rotation and thus the swimming behavior of the cell. The CheZ phosphatase speeds up the dephosphorylation of phospho-CheY reducing its concentration. This allows the flagellar rotation to return to normal and restores the cell's original swimming behavior. Abbreviations: RBP, ribose binding protein; P, phosphate group; SAM, S- adenosylmethionine; A, CheA; W, CheW; Y, CheY; R, CheR; B, CheB; Z, CheZ.


Aspartate
(Tar)
Dipeptide
(Tap)
Serine
(Tsr)
Ribose
(Trg)
RBP
CH3
R
SAM
CH3OH
ATP
ADP
A
B
P
FAD
Aer
Y
Z

primarily in chemotaxis receptors, this family of receptors has been termed methyl-accepting chemotaxis proteins or MCPs.

The methylation of a receptor can also function as a cell's memory. If the attractant level has been high recently, the methylation levels will be high. If the attractant levels were low, the methylation levels will be low. The same holds true for repellents, but in a reverse orientation. Thus, the cell can compare the current concentration gradient to one in the past based on the level of receptor methylation and can regulate the histidine kinase activity to stimulate/inhibit tumbling frequency accordingly (Blair, 1995; Stock and Surette, 1996).

Between the K1 and R1 methylation regions lies the HCD domain. Sequence analysis of the HCD regions in the four *E. coli* receptors revealed that they have a sequence similarity greater than 85%. Also, homologous domains were found in other MCPs from different bacterial species, including archea. This suggests that the HCD domain is part of a widespread motif important to receptor function (Le Moual and Koshland, 1996). This region may be so highly conserved since it is here that CheA and the docking protein CheW form a complex with the receptor (Bourret et al., 1993). The binding of an attractant sends a signal down the α4/TM2 through the CheW docking proteins to CheA (Ninfa et al., 1991; Borkovich et al., 1989).

Once the signal is received, CheA autophosphorylates and transfers the phosphate group to another response regulator, CheY. CheY is bound to CheA until the phosphate group is transferred generating phospho-CheY (CheY~P) (Welch et al., 1998; Swanson et al., 1993). Since CheY~P has a lower affinity for CheA than CheY alone

(Swanson et al., 1993), the CheY~P disassociates and diffuses into the cytoplasm where it comes into contact with the flagellar motor (Welch et al., 1993). CheY~P binds to the FliM protein in the flagellar motor switching the flagella rotation to a CW direction causing the cell to tumble (Barak and Eisenbach, 1992) (Figure 5). In its native state, CheY does not affect the flagellar rotation (CCW rotation) allowing the cells to swim in a smooth manner (Stock and Surette, 1996).

Since the half-life of CheY~P is about 10 sec (Hess et al., 1988; Stock et al., 1991) and cells need to respond to concentration changes at a much quicker rate, there needs to be a way to quickly return CheY to the unphosphorylated state. This is the role of the phosphatase CheZ (Stock and Stock, 1987). CheZ has a high affinity for CheY~P, and will rapidly cleave the phosphate from CheY (Blat and Eisenbach, 1996) allowing the flagella rotation to return to a CCW state (Kuo and Koshland, 1987). Unlike CheY, CheZ will not dephosphorylate CheB (Hess et al., 1988). CheB~P is unstable and will rapidly undergo autocatalytic dephosphorylation. Therefore it is not dependent on a separate phosphatase (Stewart, 1993).

B. Bioinformatics

Watson and Crick discovered the structure of DNA in 1953 signifying a new era in the field of molecular biology (Watson and Crick, 1953). Just seven years before that, in 1946, the first computer called ENIAC was built signaling the start of the computer age (Kidwell and Ceruzzi, 1999). Since then our knowledge in both areas has grown at a phenomenal rate. It seems only natural that these two fields would merge together into a new field of study termed 'bioinformatics' - a term that did not appear until around 1991

(Franklin, 1991). The field of bioinformatics incorporates a combination of molecular biology and computers to analyze and interpret data, develop new algorithms and statistical analysis and maintain and manage data collections (Boguski, 1998). Its roots can be traced back to the early 1960s when people like Russel Doolittle (Doolittle, 1997), Margaret Dayhoff (Dayhoff, 1969) and Walter Fitch (Fitch and Margoliash, 1967; Fitch, 1966) began to create algorithms, maintain databases and analyze sequence data before it was commonplace.

One aspect that has grown out of the bioinformatic field is the creation of sequence databases. In the early 1970s sequence collections consisted primarily of RNA sequences from a variety of microorganisms and viruses (Barrell and Clarck, 1974). It wasn't until 1977 that the first mammalian mRNA sequence from rabbit β-globin was cloned. (Baralle, 1977; Efstratiadis et al., 1977; Proudfoot, 1977). In 1982 sequence databases entered the automation era with the creation of GenBank. Here sequences were stored on magnetic tape and distributed quarterly to academic institutions (Smith, 1990). Advances in sequencing techniques have greatly increased the amount of sequences available from approximately 2,000 in the early 1980s to more than 2.7 million deposited today (Benson et al., 1999).

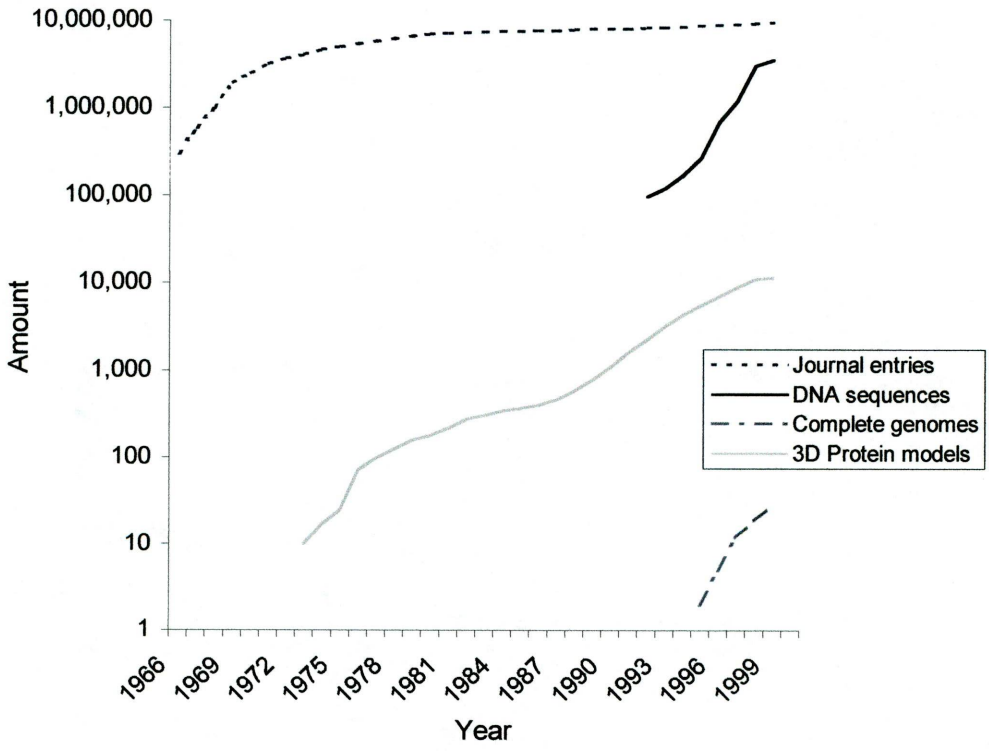
Use of new bioinformatic programs in the analysis of these sequences has lead to the discovery of new protein motifs and uncovered unexpected relationships between apparently unrelated proteins. Searches to identify PAS domains performed in 1997 identified approximately 50 domains (Zhulin et al., 1997). Now, due to advances in bioinformatic tools, such as PSI-BLAST (Altschul et al., 1997), over 300 PAS domains

have been identified (Taylor and Zhulin, 1999). Researchers have also identified SH3 domains in prokaryotes (Whisstock and Lesk, 1999) including an SH3-like domain in the CheA protein of *Thermotoga maritima* (Bilwes et al., 1999). SH3 domains play an important role in both cell-cell communication and signal transduction, and had previously been identified only in eukaryotes and viruses (Musacchio et al., 1994).

Bioinformatics has also jumpstarted the field of comparative genomics. The first completely sequenced genome from *Haemophilus influenzae* was completed and released in 1995 (Fleischmann et al., 1995). Now, approximately four years later, there are over twenty completely sequenced genomes with an additional eighty scheduled for completion by the early 2000s (http://www.tigr.org/tdb) (Figure 6) and the completion of the human genome by the year 2003. These completely sequenced genomes are influencing every aspect of modern biology from the search for new drug targets to identifying gene function (Koonin, 1999).

With so much information available, the propensity for errors remains great. Flowing from the massive amount of sequence data, both in complete genomes and in databases like GenBank, the major focus has shifted to the annotation of individual genes (Thornton, 1998). By comparing unknown proteins with those of known function, functional annotation is assigned. This annotation is based largely on homology between unknown and known proteins and is the main entry point for errors. The relationship between protein structure and function is not always a straightforward one. Just because a protein may be homologous with another doesn't necessarily mean that they have the same function. During evolution the protein's function could have changed. Similarly,

Figure 6. Graphical representation of growth in biological information. Journal entries accessible via PubMed (dotted line) located at the National Center for Biotechnology Information (NCBI) total over 9.5 million. Total number of DNA sequences available through Genbank (solid black line) is roughly 3.5 million. While there are more journal entries than DNA sequences, it is interesting to note that the total number of DNA sequences greatly outnumbers descriptive journal records. This represents a huge gap between identification of a sequence and knowledge of its function. The number of 3D protein models (solid gray line) deposited in the Protein Data Bank is >9,000. Complete genome sequencing (dashed line) is a relatively new field. Currently there are 21 published completely sequenced genome with over 50 more in the pipeline for completion by the year 2004. Adopted from (Boguski, 1998).


10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
Amount
1966
1969
1972
1975
1978
1981
1984
1987
1990
1993
1996
1999
Year
Journal entries
DNA sequences
Complete genomes
3D Protein models

two unrelated proteins may perform the same or similar functions (Thornton, 1998). The outcome is that functional annotation of unknown proteins can be wrong even if the homology is high.

This problem is of particular importance when dealing with the annotation of completely sequenced genomes, with a majority of gene function assignment based on homology and not laboratory experiments. In a recent comparison of *Mycoplasma genitalium* genome annotation by three different groups revealed an error rate of at least 8%. This means that out of the 468 reported genes in *M. genitalium*, the function of at least 40 genes do not agree. Thus, some genes had the same function assigned to them by all three groups, while other genes were assigned three different functions (a different one per group) (Brenner, 1999). This comparison illustrates how easily errors can occur and how common they are today.

When faced with the chance of error, it is important to remember that the use of bioinformatics as a research tool should not replace traditional experimental biology, but rather complement it. Knowledge gained from bioinformatic analysis can predict functional information about a gene. From that prediction a hypothesis can be developed and experiments performed to verify or disprove it (Boguski, 1998; Thornton, 1998). This is especially important in the pharmaceutical field where time and money are major considerations. Using bioinformatic tools, high quality targets for drug design can be identified and experiments performed to test the potential of these drugs.

C. Purpose of this Research

The purpose of this research was to compile a list of the most common and useful bioinformatic tools and utilize them to analyze particular aspects of the bacterial chemotaxis pathway. For my research I focused on two main components of the pathway: the CheY protein and the receptors. In this work I report on my analysis of all identified CheY proteins and CheY-like domains from different species focusing on identification of conserved functional residues and function predictions for CheY homologs and domains. Previous studies have analyzed a few chemoreceptors receptors (Le Moual and Koshland, 1996), however no recent study on newly identified receptors has been done. In this work, I also present my analysis of chemotaxis receptors including topological studies and classification, N-terminal and C-terminal analysis and predictions of transducer origins.

CHAPTER TWO
COMPILATION OF COMPUTER PROGRAMS

A. Introduction

With the advent of the Internet, various bioinformatic tools that were once only available to a select few have become available to everyone who has a computer and Internet access. In the course of my research, a wide variety of computer programs were utilized to perform different functions. While useful, there was no central list of the locations or descriptions of these various programs. One of the aims of the research performed was to compile a list of all of the various programs taking into account the type of research currently being performed in our lab at Loma Linda University. Included in this compilation would be the locations of the programs, computer system requirements (if any), a complete description of the programs and explanations of what the user can achieve by using the programs.

The compilation is presented on the department's official webpage. The webpage is divided into four different sections. The first section describes the various resource pages already compiled and accessible via the Internet. These resource pages can be divided into three main classes: 1) Comprehensive, includes those created and maintained by curators of non-redundant databases; 2) Comprehensive-hyperlinked, which compile links to various programs and 3) Specialized, those pages that detail one particular field, usually focusing on genomics.

B. Resource Pages

1. National Center for Biotechnology Information (NCBI) - Comprehensive

   Location: http://www.ncbi.nlm.nih.gov/index.html

   Description: The NCBI homepage offers access to a variety of programs. From the main page one can access PubMed (a searchable archive of journal abstracts in Medline, hyperlinked with the full-length articles at journal sites); BLAST (the NCBI's sequence similarities search algorithms); ENTREZ (the NCBI's protein/DNA sequence database search program); BankIt (NCBI sequence submission form); OMIM (Online Mendelian Inheritance in Man – a catalog of human genes); Taxonomy (searchable taxonomy browser) and Structure (database of 3D structures). The user can also access ORF finder (see page 38), which allows one to determine potential open reading frames.

   NCBI allows the user access to unfinished microbial genomes. The user can search these genomes via the BLAST programs; however it is important to note that these microbial genome sequences are unfinished and still contain errors. There may be misassembled sequences, which may result in contigs that are not accurate reflections of the finished sequence. Thus care and caution must be used when using any sequence retrieved. The sequences in these databases are presently not available at NCBI or GenBank and must be retrieved by the users from the sequencing center's FTP or website.

   Besides these main programs there are links to research projects currently underway at NCBI. Included are programs involving the human genome

sequencing, mouse/human comparisons and cancer gene research. Provided at the top of the page are links to the two major governmental agencies that support the NCBI and its work: the National Library of Medicine (NLM) and the National Institutes of Health (NIH).

2. Molecular Biology Resources at the University of Lincoln-Nebraska (CMS-SDSC) - Comprehensive-Hyperlinked

   Location: http://www.unl.edu/stc-95/ResTools/cmshp.html

   Description: This website represents a compilation of various Internet based programs. The links are divided into eight different categories including: 1) Protein Analysis and Biochemistry, 2) DNA Analysis and Molecular Biology, 3) Biomolecular Modeling, 4) Bioinformatics and Computational Biology, 5) Phylogeny and Molecular Evolution, 6) General Biochemistry, 7) General BioScience Resources and 8) Biotechnology.

   Each category is further subdivided into groups of links that perform specific functions. For example in the Protein Analysis and Biochemistry category there are sub-divisions that list programs that can be specifically used to analyze protein motifs, sequence homology, transmembrane motifs or sequence alignments. Overall, this page contains links to a majority of the most commonly used programs available, and is an excellent page to start from when searching for bioinformatic tools.

3. The Institute for Genomic Research (TIGR) - Specialized

   Location: http://www.tigr.org

Description: TIGR is a private, non-profit company whose research is devoted to genomic sequencing and structural/functional analysis of eukaryotic (plant and animal), prokaryotic and viral genomes.

One of the main features of the TIGR site is the TIGR Microbial database (TDB). This database represents a complete list of all publicly funded genome projects, finished and in progress. All of the completed genomes are searchable by either sequence similarity (BLAST-like) or by putative gene function identification (ENTREZ-like). TIGR itself has currently sequenced the following microbes: *Archaeoglobus fulgidus, Borrelia burgdorferi, Haemophilus influenzae, Helicobacter pylori, Methanococcus jannaschii, Mycoplasma genitalium, Plasmodium falciparum* (Chromosome 2) and *Treponema pallidum*. Those genomes not sequenced by TIGR or those that are not finished are still searchable via their own individual webpages. Currently there are 21 completely sequenced genomes. At the time of writing, over 80 more genomes are currently being sequenced, of which 20 have an anticipated completion date by the year 2000.

The rest of the webpage sections deal with the programs that can analyze DNA or protein sequences and those that can be utilized in phylogenetic studies. Many of the programs discussed below have specific requirements regarding the inputting of sequence data. Many require the use of certain residue codes for both nucleic acids and proteins. Some programs also require that the input sequences be in a certain format. See Appendix A for complete details.

C. DNA Sequence Analysis

A. Sequence manipulation tools

1. Readseq

Location: http://dot.imgen.bcm.tmc.edu:9331/seq-util/Options/readseq.html

Computer requirements: Any computer with an Internet connection.

Description: This webpage based program allows the user to remove amino acid numbers and/or spaces from DNA sequences that have been obtained via ENTREZ or BLAST searches. The output file will be in standard FASTA format. It can process multiple sequences at one time, provided that they are all in the correct format. This program also supports a wide variety of input formats, including some of the more common ones like PHYLIP, PAUP/Nexus, FASTA, GenBank (GB) and Fitch.

Purpose of program: This program is designed to save the user time by arranging a sequence into FASTA format. This is especially useful if one has a long sequence that needs to be converted into the correct format.

2. Reverse Complement

Location: http://dot.imgen.bcm.tmc.edu:9331/seq-util/Options/revcomp.html

Computer requirements: Any computer with an Internet connection.

Description: Sequence orientation conversion. With this program, one can reverse, complement, or reverse and complement a DNA sequence. Like

Readseq, this program can support multiple sequences at once provided that they are in the same format.

Purpose of program: This program makes analysis of DNA sequences easier. The user can simply input his sequencing data and the program will reverse or complement it.

3. Translate tool

Location: http://www.expasy.ch/tools/dna.html

Computer requirements: Any computer with an Internet connection.

Description: This is a tool that allows the translation of a nucleotide sequence (DNA/RNA) to a protein sequence.

Purpose of program: With this program, a nucleotide sequences can be converted into an amino acid sequence. It takes three nucleotides to code for one amino acid so there are three different open reading frames (ORFs) in the 5′ to 3′ direction that can encode a protein of interest. Since DNA is complementary, there are also three reading frames that can encode a protein in the 3′ to 5′ direction. Unlike many of the other programs accessible via the Internet, translate tool will translate a sequence in all six different reading frames.

The other important feature of this program is that it will allow the user to create a virtual Swiss-Prot entry comprised of the query residues from one of the ORFs. From this virtual entry, a sequence can be directly submitted to various other Internet tools including BLAST (at the NCBI), SWISS-Model,

ProtParam and ProtScale. This allows one to quickly search for any possible motifs or domains that the sequence might contain as well as search for any homologous proteins in the various databases.

B. Primer design, PCR and restriction mapping

1. Primers!

Location: http://www.williamstone.com/primers/javascript/

Computer requirements: Any computer with an Internet connection and a web browser that supports JavaScript (Netscape Communicator or Internet Explorer 4.0+).

Description: By inputting a DNA sequence, forward and reverse primers can be created based on specified properties like T_m (temperature at which half of the primer ends are annealed) and primer length. The user may also restrict primers to specified locations in their DNA sequence (5′ to 3′ or 3′ to 5′ forward/reverse primers).

The output is displayed in a table format listing the primer sequences, T_m and location of the primers in the original sequence. From this table, one can select the forward and reverse primer that best fits the user's requirements. An analysis of both the input DNA sequence and the chosen primers is also performed. This analysis graphically portrays the location of the primers on the input sequence, the T_m calculated by the various methods and any possible problems with the primers. The three main problems that are address include: 1) Hairpin analysis - where primers can fold and anneal

with themselves disrupting binding with the target DNA, 2) Primer-Dimer analysis - where one copy of a primer may anneal with other copies of themselves disrupting DNA binding and 3) Primer Pair similarity - where the forward and reverse primers may bind with each other also disrupting binding to the target DNA. Finally the webpage has the option to let the user order the derived primers on-line.

Purpose of program: This program was designed to simplify the process of primer creation for PCR and DNA sequencing. No longer does one have to scour a DNA sequence by eye and calculate the T_m by hand. All of these parameters are taken into account in the primer generation. Its analysis of the common problems associated with primer formation is also helpful.

However, this program does lack some features. It doesn't take into account any 3' GC clamps added to the primer sequence. GC clamps are used to help bind and stabilize the primer binding to the target DNA. It also doesn't generate degenerate primers, which can be used to mutate one amino acid into another via changes introduced in the nucleotide sequence.

2. Primer Design

Location: http://www.embl-heidelberg.de/~toldo/JaMBW/5/2/index.html

Computer requirements: Any computer with an Internet connection and a web browser that supports JavaScript (Netscape Communicator or Internet Explorer 4.0+).

Description: A more simplified primer designed program. It includes all of the standard parameters (T_m, reagent concentrations, primer GC percent and primer length). Unlike Primers!, this program takes into account the number of GC clamps wished in primer construction.

Purpose of program: This program analyzes DNA sequences and generates primers that can be used for PCR or sequencing reactions. It has many of the same functions that Primers! has with the addition of a few new parameters including GC clamps.

3. T_m determination

Location: http://alces.med.umn.edu/rawtm.html

Computer requirements: Any computer with an Internet connection.

Description: This program will determine the T_m of a DNA sequence that will be used as a primer for PCR reactions. This program assumes that the input sequence will not be symmetric and will contain at least one G or C. The minimum length for this program is eight nucleotides. The salt (mM) and DNA (nM) concentrations may also be adjusted.

4. Webcutter

Location: http://www.medkem.gu.se/cutter/

Computer requirements: Any computer with an Internet connection.

Description: Webcutter in an on-line tool that can be used to identify various restriction sites in a DNA sequence. The user inputs a sequence and selects from a variety of options including the type of sequence analysis wished,

which enzymes to use in the analysis and how the output will be arranged. The output will arranged in the following sections: 1) a graphical representation of the input sequence with all of the identified enzymes listed at their respective cutting sites, 2) an alphabetical list of the restriction enzymes, the location (nucleotide number) where they cut and their recognition sequence and 3) a list of all the enzymes in the database that do not cut the input sequence.

Purpose of program: By utilizing this program, restriction enzymes for cloning or digestion of a DNA sequence can be quickly identified. Webcutter does have some limitations. Currently the program only supports type 2 restriction enzymes which cut at a specific nucleotide location once. These are most common type of enzymes used in most experiments. There are two other types: type 1 and type 3. These cut differently, are not used as often and thus not included in the Webcutter database.

C. Structure and Function Prediction

1. ProScan

Location: http://bimas.dcrt.nih.gov/molbio/proscan/index.html

Computer requirements: Any computer with an Internet connection.

Description: Promoter Scan (ProScan) is used to identify putative eukaryotic Polymerase II (Pol II) promoter sequences. If it identifies a putative promoter site, the program will also try to identify the TATA box. Based on TATA box location the program estimates a transcription start site.

Purpose of program: This program is designed to help identify putative eukaryotic promoter regions. Many times Pol II promoter sites lie upstream of the sequence they are going to translate. These sites are usually spread out over a region >200 base pairs and the sequences between them are not always important. Thus it is crucial that the correct promoter site be identified. ProScan performs this task via computer algorithms that recognizes of approximately 70% of primate promoter sequences, with a false positive rate of about one in every 14,000 bases.

2. Gene Finder

Location: http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: Like ProScan, this program is used to analyze eukaryotic DNA sequences to identify splice sites, protein coding exons, promoter regions and poly-A tail signals. Unlike ProScan, Gene Finder compares the input sequence with various databases. The user has the option to select from up to five different organism databases as well as the option to select various search algorithms.

3. FramePlot

Location: http://www.nih.go.jp/~jun/cgi-bin/frameplot.pl

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: Predicts protein coding regions in all ORFs in bacteria especially those with a high G+C content (Ishikawa and Hotta,

1999). There are many options available including selection of start codon, minimum ORF size and incomplete ORF recognition.

The output is a graphical representation of the query sequence. On the top, in various colors, are lines representing the putative encoded proteins in the different ORFs. By clicking on an individual line, the sequence it represents is displayed in FASTA format. The program is connected with BLAST; thus the putative protein can be directly submitted for analysis.

4. ORF Finder

Location: http://www.ncbi.nlm.nih.gov/gorf/gorf.html

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: The Open Reading Frame Finder (ORF Finder) is a graphical analysis tool which finds all open reading frames of a user's sequence. The deduced amino acid sequence can be saved in various formats and searched against the sequence database using the BLAST server.

5. GeneMark

Location: http://genemark.biology.gatech.edu/GeneMark/hmmchoice.html

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: The GeneMark server provides e-mail based identification of protein coding regions in DNA sequences from prokaryotic and eukaryotic species. The GeneMark algorithm was designed to improve the gene prediction quality in terms of finding exact gene boundaries using the hidden Markov model framework and potential

ribosome binding sites (RBS) (Lukashin and Borodovsky, 1998). The GeneMark server accepts a formatted message containing a DNA sequence in text format. You may specify a species name and parameters that control the analysis procedure. The server will send the user a graphical representation of the input DNA. The display includes all six ORFs in which solid black lines represent the most likely genes. Any other sections that may represent genes are shaded gray.

D. Sequence retrieval and database searches

1. BLAST

Location: http://www.ncbi.nlm.nih.gov/BLAST/

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: The Basic Local Alignment Search Tool (BLAST) is a similarity search program designed to analyze all available DNA/protein sequence databases (Altschul et al., 1990). Table 2 compares the different types of BLAST searches and their common uses. As one can see, BLAST searches at the DNA level are usually only performed when one wants to identify identical DNA sequences and splicing sites. The most useful information that a user can obtain from BLAST searches comes at the protein level, so a complete description of the program can be found under the protein sequence analysis section (see page 45)

2. ENTREZ (DNA searches)

Location: http://www.ncbi.nlm.nih.gov/Entrez/nucleotide.html

Table 2. List of the variant BLAST search algorithms[a].

Program	Query	Database	Comparison	Common use
blastn	DNA	DNA	DNA level	Identify identical DNA sequences and splicing sites
blastp	Protein	Protein	Protein level	Find homologous proteins
blastx	DNA	Protein	Protein level	Converts DNA to protein sequence to find both potential genes and seek homologous proteins
tblastn	Protein	DNA	Protein level	Searches for genes in unannotated DNA and unfinished genomes
tblastx	DNA	DNA	Protein level	Discover gene structure

[a]DNA sequences converted to protein are searched in all six potential open-reading frames.

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: A text-based molecular biology database search program. The above link is specifically for retrieval of DNA sequences. For a complete description of ENTREZ see below.

D. Protein sequence analysis

A. Sequence retrieval and database searches

1. ENTREZ (Protein searches)

Location: http://www.ncbi.nlm.nih.gov/Entrez/protein.html

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: ENTREZ is a molecular biology database search and retrieval system. From this location, a wide variety of integrated databases can be explored. The user can specify which of the supported databases are to be searched. Currently the databases that ENTREZ supports are GenBank, EMBL, DDJB, SWISS-PROT, PIR, PRF, PDB (both protein sequence and 3D structures) as well as PubMed (which contains over 9 million journal entries from Medline).

The purpose of ENTREZ is to allow the user to retrieve DNA/protein sequences based on a number of different criteria including (but not limited to) gene name, accession number, author's name, journal name and function. Upon a query, ENTREZ will search all desired databases and display the results. Depending on the number of hits, the user may wish to focus the search by adding more terms to the query (i.e. adding the species name that

contains the gene of interest). By adding more specific terms the user will eventually obtain a small list of hits that pertain to the original query.

From the list, each hit may be viewed in GenBank report, FASTA report, ASN.1 report or Graphical view. ASN.1 is a programming language that is used at the NCBI. All of the information is contained there, but in a computer language format. The Graphical view simply shows the sequence in a graphical representation of its location in the genome. This is particularly helpful if the gene is part of an operon. By selecting the Graphical view, one can determine where in the operon the gene is located and what genes may surround it.

Of the four different views, the GenBank and FASTA are the most useful. The FASTA report displays the sequence in the FASTA format that can then be used for further analysis. The GenBank report is the most detailed report. It provides the accession number, species name, where the sequence is published and the authors, annotation of possible function and displays the sequence including residue numbers.

There are some minor differences between the GenBank report of a DNA and protein sequence. The DNA report will allow access to the protein report and visa versa. However, the protein report will also contain a link that will pull up related sequences. This is helpful in retrieving related sequences that may have been missed in the original query. Sometime there will be the option at the top of the page to visit the sequence's literature

report either in PubMed or at the journal's webpage. From there the user can gain access to the entire article and download or print it for a few dollars.

There are a few potential problems that users of ENTREZ should be aware of. First, since all of the sequences are submitted and annotated by hand, there is the danger of mislabeled or wrongly identified proteins. For example, there is a gene from *Bacillus subtilis* that was deposited under the name of CheB (accession number 142682). However, further experimentation determined that this gene behaved more like the CheY protein. Based on this research this protein is now referred to as CheY not CheB, yet the original entry still exists. A new entry has also been created to reflect the name change (accession number 116278). As a result there is an incorrectly annotated CheB gene and a correctly annotated CheY gene in the database.

Also with the recent focus on sequencing entire genomes, multiple entries of the same gene can be found. Each entry contains the same sequence yet it has been deposited under a different accession number and different authorship. As a result, there can be quite a bit of redundancy in the database. A careful study of identical gene hits from the same organism should help eliminate any repetition.

Care should also be taken when performing ENTREZ searches to help eliminate potentially misleading hits. The user should try narrowing a search at the beginning by selecting the most appropriate search field. The default setting is "All Fields." By changing the search fields, the user may be able to

reduce the number of misleading hits. For example, a search using the default settings for the chemotaxis protein CheW results in 137 hits. A majority of those hits do not relate to chemotaxis. A closer look reveals that a majority of the hits contain authors whose last name is Chew. By changing the search field to "gene name" the number of hits is reduced to 26 actual CheW genes.

2. SWISS-PROT

Location: http://expasy.hcuge.ch/sprot/

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: SWISS-PROT is an annotated protein database maintained by the University of Geneva. Unlike many other databases SWISS-PROT prides itself on having the most comprehensive annotation possible. Annotation can consist of the following: protein function, post-transcriptional modification, domains and various other sites (i.e. ATP-Binding sites, zinc fingers, etc.), secondary and quaternary structures, similarity to other proteins and any variants or sequence conflicts with other proteins.

SWISS-PROT also takes great steps to insure that there is a minimal amount of redundancy. They do this by combining all of the data entries into one. Then, if there are any variations they are reported in the annotation.

Thirdly SWISS-PROT is also currently integrated with thirty other databases (similar to ENTREZ). This allows the user to quickly compare the results generated by a SWISS-PROT search with those from these other

databases. Currently there are 78,841 individual entries in the SWISS-PROT database.

3. BLAST

Location: http://www.ncbi.nlm.nih.gov/BLAST/

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: As stated earlier, BLAST is a similarity search program designed to analyze different sequence databases. BLAST has been designed for speed and uses a heuristic algorithm to detect regions of similarity among different sequences.

In order to determine statistical significance of a similarity rather from that of a random hit, BLAST makes use of the Expected value (E-value). The E-value is defined as the number of hits that can be expected due to random chance when searching a database (Altschul, 1998). Thus the lower the resulting E-value, the more likely the value is due to homology and not random chance.

Another area that needs to be addressed is the low-complexity region. These are regions in the DNA/protein that contain highly biased residue composition and may contain many short repeats (Wootton and Federhen, 1996). These low-complexity regions can create problems in similarity searches by retrieving many sequences that are not truly homologous. BLAST employs certain filters that help eliminate low-complexity regions from similarity analysis. However, care must be taken in the interpretation of

results to make sure that homology isn't due to any low-complexity regions missed by the filters.

To perform BLAST searches, the user must input a query sequence (DNA or protein) in FASTA format. The results are displayed as a clickable picture showing the different protein's regions of similarity in comparison to the query. Below the picture is a list in descending order of the most significant hits based on E-value. Besides E-value the accession number, protein name and Score bit are also listed. By clicking the Score bit, an alignment of the regions of homology between the protein and the query are shown. Here the percent identity (number of residues that are similar between the two sequences), percent positives (number of residues that are identical between the two sequences) and the percent of gaps (number of gaps introduced between the two aligned sequences) are displayed. Since BLAST is integrated with ENTREZ (see page 41), clicking on the accession number brings up the database entry for that particular protein. Residue series of NNNNNNNNN (nucleotide sequences) or XXXXXXXXXX (amino acid sequences) represent areas of low-complexity that have been filtered.

By performing BLAST searches, proteins that contain homologous regions can be identified. This allows the user to identify common motifs shared between the two proteins. From this information a hypothesis can be made as to possible function and structural importance of the motif. BLAST searches are also helpful in determining possible gene function of newly

sequenced DNA (Table 2). By comparing the function of similar proteins users can predict possible functions and design experiments to prove or disprove the prediction.

Interpretation of BLAST search results is a problematic procedure. It is dangerous to place complete trust in the results of a search without a more through examination. One concern is the possibility that some genes might contain homologous regions, yet overall have entirely different functions. This would indicate that perhaps the region identified isn't necessarily significant, i.e. a low-complexity region that was not filtered.

Another problem lies in the fact many times the areas that contain the best (lowest) E-values are contained in one region of the protein. People have a tendency to focus only on those hits that have low E-values. There may be other equally significant regions that have a higher E-value and be further down on the results list. The danger is that important motifs may be missed. It is also important to remember that just because there are no hits doesn't preclude the possibility that there are still homologous proteins that have been sequenced. BLAST doesn't search every protein database in existence. A thorough search of many different databases might result in significant hits.

4. PSI-BLAST

Location: http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: Position-Specific Iterated BLAST (PSI-BLAST) is a derivative of BLAST that performs iterative database searches and can uncover many protein relationships that are not detectable via one-pass database searches (Altschul et al., 1997). PSI-BLAST works as follows:

1) A query sequence is inputted and compared with the various databases available using the normal BLAST search algorithm (Altschul et al., 1990). From this search, a multiple alignment is constructed using the query as the template. The alignment forms the basis for a profile of the aligned regions (Altschul et al., 1997).
2) PSI-BLAST then compares this profile against the databases again trying to find any other regions that can align. Once the profile has been compared again, statistical significance of the resulting hits is estimated. For statistical analysis, PSI-BLAST uses E-values to determine significance (Altschul et al., 1997).
3) Finally, PSI-BLAST can iterate (repeat) searches by creating a new multiple alignment including all new hits and re-searching the database. This process can be repeated any number of times until user intervention or the program reaches convergence. Convergence occurs when no new statistically significant sequences can be detected.

Like BLAST, the lower the E-value the more significant the sequence alignment. By default the cutoff E-value for PSI-BLAST is .001. The user has

the option the change the E-value to whatever they feel is appropriate. However, care must be taken. If the E-value is put too high, sequences might be included in the alignment that may not necessarily be significant.

E-values are not the only option that the user may change. PSI-BLAST displays the results of a search in a similar fashion to the regular BLAST output. The major addition is a separation of the resulting sequences based on their E-values into two groups: those above the E-value threshold and those below. Only those sequences with E-values above the threshold are automatically included in the next iteration. However, the user may manually select any sequence to be included in the next iteration regardless of E-value.

Once a single sequence from a highly conserved family is used in constructing a profile, the rest of the family will almost certainly be retrieved (and have E-values of high significance) in subsequent iterations. Impressive E-values for sequences retrieved in later iterations depend upon the validity of earlier inferences and therefore should not be taken as automatic proof of homology.

PSI-BLAST is a powerful tool, but its use requires caution. The sources of error are the same as for standard BLAST but are easily amplified by iteration. The major source of deceptive alignments is the presence within proteins of regions with highly biased amino acid composition (Wootton and Federhen, 1996). If such a region is included in a profile production of

otherwise unrelated sequences containing similarly, biased regions will probably creep in during subsequent iterations rendering the search nearly worthless.

PSI-BLAST filters out biased regions of query sequences (like low-complexity regions) by default. Because the parameters have been set to avoid masking potentially important regions, some bias may persist; PSI-BLAST can thus still generate compositionally rooted artifacts. These cases usually can be identified by inspection - especially when sequences have a known bias (i.e. myosin or collagen).

5. BLAST 2 sequences

Location: http://www.ncbi.nlm.nih.gov/gorf/bl2.html

Computer requirements: Any computer with an Internet connection.

Description/Purpose of Program: This tool produces the alignment of any two sequences using the BLAST engine for local alignment (Tatusova and Madden, 1999). The output is represented in a graphical representation, with the areas that over lap represented by thick blue lines. The actual residue sequences can also be viewed. This program is helpful for a user trying to align multiple DNA sequences that have been obtained by primer extension or for checking site-directed mutagenesis work.

B. Multiple Alignment

1. Clustal W

Location: http://bioweb.pasteur.fr/seqanal/interfaces/clustalw.html
For color output: http://www2.ebi.ac.uk/clustalw

Computer requirements: Any computer with an Internet connection.

Description: This algorithm generates multiple alignments of protein/DNA sequences (Thompson et al., 1994). It also has the option of generating a phylogenetic tree based on the Neighbor-Joining method. Unlike Clustal X (see below), this program is available only via the Internet. Since it is run on a Silicon Graphics computer, the number of input sequences it can align is rather high. Clustal W also provides the alignment in FATSA format, which can be important for further analysis.

2. Clustal X

Location: ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalX

Computer requirements: Clustal X is supported on a variety of different platforms including: Microsoft Windows (32 bit) for PCs, SUN Solaris, IRIX5.3 on Silicon Graphics, Digital UNIX on DECstations, Linux ELF for x86 PCs, and Macintosh PowerMac.

Description/Purpose of Program: Clustal X is a downloadable program that can generate multiple alignments of both protein and DNA sequences (Jeanmougin et al., 1998; Thompson et al., 1997). Sequences in the FASTA format are loaded into the program and the output is displayed in a graphical representation, with the various amino acids color-coded. Included in the alignment is a representation of the percent similarity of each aligned residue portrayed as a plotted line drawn under the complete alignment. The higher the peaks on the line the higher the percent similarity.

Clustal X also has the unique feature that allows the modification of the generated alignment by rearranging the order of the input sequences without disturbing the alignment itself. For example the user can cut and paste sequences from a specific bacterium so that they are all clustered together. This feature also allows for the deletion of a sequence from the alignment. Thus if the sequence is inputted more than once by accident or with a different name, the user can delete the duplicates.

Another modification feature supported by Clustal X is the ability to select a specific residue range and rearrange only those residues. If a user feels that a certain range of residues are not properly aligned, simply select the range of residues to be looked at, highlight the "rearrange only selected residues" option and those residues will be aligned. Clustal X will not touch any residue that falls out side the range, so the rest of the alignment will be the same. The user should note that the program will ask for new output file names each time a rearrangement is done. If the user wishes to keep the original alignment as a backup, then new file names should be assigned. Also note that there can not be any spaces in the folder name where the output file is to be saved, otherwise an error message will occur and the file won't be saved.

Based on the generated alignments, evolutionary relationships between the various sequences may be analyzed. Clustal X can perform this analysis by generating phylogenetic trees. The method used to generate these

trees is the Neighbor-Joining method (NJ method) first described by Saitou and Nei (1987). Using this method the percent divergence (distances) between all of the sequences in the multiple alignment is calculated. Next the NJ method is applied to this distance matrix resulting in the tree generation.

There are two options for tree generation in Clustal X. The first is the standard NJ tree and the other is a bootstrapped NJ tree. Bootstrapping is a statistical method for determining confidence values for the branch points (groupings) on the tree. Bootstrapping involves taking a random number of random samples of sites from the alignment (N - should be a large number), drawing N trees (1 from each sample) and counting how many times each grouping from the original tree occurs in the sample trees. The higher the number at the branch, the more confident that it is correct.

Note: The draw tree option in Clustal X produces an unrooted tree. In order to determine the root of a tree, an outgroup (a sequence that you are certain branches at the outside of the tree, usually not related to the other sequences) must be selected. The easiest way to visualize and select an outgroup is to use a third-party viewer program like TreeView (see page 63).

Finally, Clustal X can generate and save the alignments in a variety of different formats including CLUSTAL, GCG, NBRF/PIR, PHYLIP and GDE. The colorized alignment can also be exported as a PostScript file for printing. Using the default color scheme included in Clustal X, this export function will allow users to print the alignment as it appears on the screen. There are a

few third-party PostScript viewers available for the IBM. The easiest to use is GhostView (http://www.cs.wisc.edu/~ ghost/gsview/).

3. Consensus

Location: http://www.bork.embl-heidelberg.de/Alignment/consensus.html

Computer requirements: Any computer with an Internet connection.

Description: Using a custom Perl Script, a consensus can be generated from a multiple alignment in either Clustal or MSF format. The user has the option to customize the threshold from 50% up to 100% via multiples of five. The groupings of the amino acids can also be modified and new groups can be added to the consensus.

Purpose of program: This program was designed to save the user time by automating the tedious task of generating a consensus. However, care should still be taken to verify that the output is correct. Since the results don't always come back perfectly aligned, the user must make sure that the final alignment and the consensus are in agreement.

C. Domain analysis and structural elements

1. DAS server

Location: http://www.biokemi.su.se/~server/DAS/

Computer requirements: Any computer with an Internet connection.

Description: The DAS server will predict transmembrane regions in a protein sequence using the Dense Alignment Surface method (Thompson et al., 1994).

Purpose of program: This program is used to determine the topology of a given protein. The program displayed the results in two formats, a list of the potential transmembrane segments and a DAS curve. The DAS curve plots the "DAS" profile score against the query sequence. These curves are obtained by pairwise comparison of the proteins in the test set against each other. There are two cutoff lines indicated on the plot. One at 2.2 represents a "strict" cutoff in terms of the number of matching segments, while the other at 1.7 represents a "loose" cutoff and gives the actual location of the transmembrane segments. The data can be used to identify different classes of membrane topology and allow the user to separate proteins into these classes.

2. ProDom

Location: http://protein.toulouse.inra.fr/prodom.html

Computer requirements: Any computer with an Internet connection.

Description: ProDom is a database that contains protein domain families identified in the SWISS-PROT database. ProDom has recently been upgraded to allow searches similar to those using an engine based on PSI-BLAST. The generated results include graphical representation of homologous domains, multiple alignments of these domains (sequenced based) and phylogenetic trees generated from the multiple alignments.

Purpose of program: ProDom is useful for identifying potential domains in a query sequence. The results are displayed as a multiple alignment and a

phylogenetic tree is generated. From this tree, hypotheses about possible domain function and evolutionary importance can be developed. One disadvantage to ProDom is that it appears to be limited only to those sequences available in the SWISS-PROT database. Thus any sequences not included in that database would not show up in the analysis. The user may want to perform additional searches for domains using other database search programs such as PSI-BLAST to make sure that all potentially important information is obtained.

D. Secondary Structure

1. PhD Server (part of the Predict Protein server)

Location: http://www.embl-heidelberg.de/predictprotein/predictprotein.html

Computer requirements: Any computer with an Internet connection.

Description: PhD is the most commonly used predictor of secondary structure (potential α-helixes, β-sheets and loops). Users submit a protein sequence to the Predict Protein server and the results are returned via email. The secondary structure is predicted via a neural network system with accuracy > 72% for the three structure types: helix, strand and loop (Rost and Sander, 1993a; Rost and Sander, 1993b; Rost and Sander, 1994). Unlike other secondary structure predictors, users can submit multiple alignments (like Clustal or MSL files) to help increase the accuracy of the prediction.

The default submission setting offers the user the option of receiving

the results in HTML format (the default being plain text). The user can then view the results in a web browser. Note: The HTML file is not normally suited for printouts. If the results are to be printed, the user must go into the advanced/expert settings and select "HTML with PHD graphs for printouts" option. Alternately the user can choose to download the results via FTP rather than email.

Purpose of program: PhD is used to predict secondary structure of protein sequences. PhD predictions have two main features: 1) improved prediction accuracy through evolutionary information from multiple sequence alignments and 2) a more accurate prediction of secondary structure segments by using a multi-level system.

2. Predator

Location: http://www.embl-heidelberg.de/predator/predator_info.html

Computer requirements: Any computer with an Internet connection.

Description: Predator is a secondary structure prediction program. The user supplies a protein sequence and Predator will identify potential α-helixes and β-sheets. The user can also supply a set of multiple unaligned sequences to help improve the secondary structure predictions. Predator does not use multiple sequence alignments, since it relies on pairwise local alignments of the sequence set included with the query sequence. If you supply a set of sequences already aligned the sequences will be treated as unaligned (Frishman and Argos, 1997; Frishman and Argos, 1996).

E. 3D protein structure

1. Swiss PDB viewer

Location: http://www.expasy.ch/spdbv/mainpage.html

Computer requirements: Currently supports the following platforms: Windows 95/98 and NT, Macintosh and soon Unix.

Description: Swiss PDB viewer is a free program that has the ability to analyze multiple protein 3D structures at the same time. Once obtained from a database, these proteins can be superimposed in order to analyze structural alignments and compare fold similarities. Visualization can be improved via the option to change secondary structure and individual amino acid colors. amino acid mutations, H-bonds, angles and distances between atoms can all be derived by various built in functions.

Swiss-Model, an automated homology modeling server running in the Geneva Biomedical Research Center is also integrated into Swiss PDB viewer. This integration makes it is possible to thread a protein primary sequence onto a 3D template, find out how well the sequence overlays the template and submit a request to generate a 3D structure of the sequence based on the overlay.

Purpose of program: Swiss PDB viewer can be used to analyze multiple crystallized proteins in the same window. Similarly SWISS-Model allows the user to model amino acid sequences with known 3D structures. These features allow the user to A) predict 3D structure of any amino acid

sequence, B) visualize any fold similarities between various proteins and C) hypothesize about the possible roles these similarities may play both in terms of domain functions and evolutionary importance.

2. RasMol

Location: http://www.umass.edu/microbio/rasmol/

Computer requirements: Currently supports the following platforms: Windows 95/98, Unix and Macintosh.

Description: RasMol is a free program that can be used to view 3D molecule images. Supported file types include Brookhaven Protein Databank (PDB), Mol2 formats, Molecular Design Limited's (MDL) file format, Minnesota Supercomputer Centre's XYZ (XMol) format and CHARMm format files. RasMol can display molecular structures as a wireframe, cylinder stick-bond, space filling (CPK) sphere, ribbon (smooth solid or wire-like), hydrogen bond or dot surface representation. Various portions of the molecule can be colored individually or displayed by themselves. The user can manipulate the displayed molecule via rotation, translation and zooming. Finally the displayed molecule can be exported in a variety of different formats including: PostScript, GIF, PPM, BMP, or PICT.

Purpose of program: RasMol is a simple program that can be used to quickly view and export 3D renderings of proteins whose structures have been determined. Concurrently, files rendered using Swiss PDB viewer can be opened using RasMol for quick viewing, modification or exportation. One of

the main limitations of RasMol is that it can not open multiple files at the same time. Thus it is not easy for the user to analyze multiple proteins for 3D homology.

3. PovRay

Location: http://www.povray.org/

Computer requirements: Currently supports the following platforms: Windows 95/98/NT, DOS, Macintosh, i86 Linux, SunOS and Amiga.

Description: The Persistence of Vision Ray Tracer (POV-Ray) is a free tool for creating 3D images.

Purpose of program: When used in conjunction with Swiss PDB viewer, the rendered output image appears much sharper and the colors are more vivid. Besides the improving of quality of the image, PovRay also offers the user more features. The user can select the resolution (i.e. 640x480, 800x600, 1024x768, etc.) and color quality (16-bit, 24-bit, 32-bit, etc.) of the rendered file. Images generated using this program can then be exported as an appropriate file type for journal publications.

E. Phylogenetic analysis

A. PHYLIP

Location: http://evolution.genetics.washington.edu/phylip.html

Computer requirements: Currently supports the following platforms: Windows 95/98/NT and Macintosh.

Description/Purpose of Program: PHYLIP (the *PHYL*ogeny *I*nference *P*ackage) is a package of different programs used for inferring phylogenies (evolutionary trees). The individual programs contained in PHYLIP include methods for calculating parsimony, distance matrix and likelihood methods (i.e. bootstrapping and consensus trees).

All of the individual programs are controlled through a menu system that allows the user to change program options and perform the computations. Data is inputted into the programs via a plain text file termed the 'infile.' If the program can not find the 'infile' the user will be prompted to type the location of the data file. Many of the programs included in PHYLIP require the sequences to be already aligned. Most alignment programs have the option to format data files in the PHYLIP format.

Below is a list of the most commonly used individual programs used in PHYLIP to generate phylogenetic trees:

PROTPARS: Estimates phylogenies from protein sequences (input using the standard one-letter code for amino acids) using the parsimony method, in a variant which counts only those nucleotide changes that change the amino acid on the assumption that silent changes are more easily accomplished.

PROTDIST: Computes a distance measure for protein sequences, using maximum likelihood estimates based on the Dayhoff PAM matrix, Kimura's 1983 approximation to it, or a model based on the genetic code plus a

constraint on changing to a different category of amino acid. The distances can then be used in the distance matrix programs.

SEQBOOT: Reads in a data set and produces multiple data sets from it by bootstrap resampling. Since most programs in the current version of the package allow processing of multiple data sets, this can be used together with the consensus tree program.

CONSENSE: Computes consensus trees by the majority-rule consensus tree method, which also allows one to easily find the strict consensus tree. Trees are input in a tree file in standard nested-parenthesis notation, which is produced by many of the tree estimation programs in the package. This program can be used as the final step for bootstrap analyses for many of the methods in the package.

DRAWGRAM: Plots rooted phylogenies, cladograms and phenograms in a wide variety of user-controllable formats. The program is interactive and allows previewing of the tree on PC graphics screens and Tektronix or DEC graphics terminals. Final output can be on a laser, graphics screens or terminals, in files readable by drawing programs and pen plotters printers capable of graphics.

DRAWTREE: Similar to DRAWGRAM but plots unrooted phylogenies.

RETREE: Reads in a tree (with branch lengths if necessary) and allows one to reroot the tree, to flip branches, to change species names and branch

lengths, and then write the result out. Can be used to convert between rooted and unrooted trees.

PHYLIP also contains a number of programs that will perform and generate phylogenetic trees for DNA sequences as well. However, it is easier to use protein sequences in analysis as it takes less computer processing power. For a complete description of the DNA programs visit the PHYLIP webpage.

Output from the various programs is saved as special files with specific names. For most of the individual programs (protein/DNA) the output file is simply called the "outfile." This is usually a simple text file that can be opened with any word processor. The data files generated by the tree programs are saved in files called "treefile". Trees saved in the "treefile" are in the Newick format, an informal standard agreed to in 1986 by authors of a number of major phylogenic packages.

B. Treeview

Location: http://taxonomy.zoology.gla.ac.uk/rod/treeview.html

Computer requirements: Currently supports the following platforms: Windows 95/98/NT and Macintosh.

Description: A simple program for displaying phylogenetic trees generated from various programs.

Purpose of program: Simplifies viewing and printing of evolutionary trees generated from different phylogenetic programs including PHYLIP.

TreeView can display a tree is four ways: 1) unrooted, 2) slanted cladogram, 3) rectangular cladogram and 4) phylogram. TreeView also has the option to display bootstrap data if bootstrapping on the tree was performed. The tree can then be saved in a native graphic format for publications or exporting into a different program. TreeView also included a basic tree editor that provides tools for modification of branch points, rerooting, polytomy formation and rearranging the appearance of the tree.

CHAPTER THREE
ANALYSIS OF THE RESPONSE REGULATOR CHEY

A. Introduction

Bacterial chemotaxis is one of the best understood signal transduction systems. It allows bacteria to sense and migrate towards optimal conditions needed for growth and survival as well as to escape harmful conditions (Hazelbauer and Adler, 1971). A bacterium's movement towards or away from these conditions is controlled by flagellar rotation (Berg and Anderson, 1973). The main components of the chemotaxis excitation pathway are: a receptor-associated histidine kinase, CheA, which can phosphorylate another component, a response regulator, CheY. In its inactive form, CheY remains bound to CheA and the flagella rotates in a default counter-clockwise rotation. Phospho-CheY (CheY~P), the active form, binds to the flagella motor switch protein, FliM, causing clockwise rotation (Berg, 1993; Silversmith and Bourret, 1999).

In order for the cell to respond quickly to stimuli changes, CheY~P needs to be quickly dephosphorylated, yet CheY~P has a half-life of about 15 sec (Silversmith et al., 1997). In vivo measurements showed that CheY~P dephosphorylation was increased by a factor of ten, greatly reducing its half-life (Silversmith et al., 1997). Hess et al. (1988) found that this was due to the phosphatase CheZ. However, a comparison of all known chemotaxis operons revealed the presence of CheZ only in γ-proteobacteria *Escherichia coli* (Mutoh and Simon, 1986), *Salmonella typhimurium* (Stock and Stock, 1987), *Pseudomonas aeruginosa* (Masduki et al., 1995) and *Pseudomonas putida* (Ditty et al., 1998). Thus other mechanisms of dephosphorylation must be involved. Some genomes contain

multiple CheY-like proteins, which may act as phosphate sinks, as demonstrated in *Rhizobium meliloti* (Sourjik and Schmitt, 1998), or the dephosphorylation of CheY~P may proceed without a phosphatase more rapidly than in *E. coli.*

Several proteins have been found to contain CheY-like domains. The best known CheY domain is located in the N-terminal receiver domain of methylesterase CheB (Hess et al., 1988; Lupas and Stock, 1989; West et al., 1995), which is a part of the adaptation pathway in *E. coli* (Stock and Surette, 1996). We have identified known and putative CheY and CheY domains through an exhaustive BLAST (Altschul et al., 1997) search of the non-redundant database using known CheY sequences as queries. The search identified CheY-like domains present in CheA from *Helicobacter pylori, Rodospirillum centenum* and *Synechocystis,* and in CheV from *H. pylori* and *Bacillus subtilis* (Table 3). These CheY domains appear to be more closely related to multiple copy CheYs than main (FliM-interacting) CheY proteins, lending credence to the phosphate sink hypothesis.

B. Domain analysis and structural identification

Based on the NMR (Moy et al., 1994) and X-ray crystallography (Volz and Matsumura, 1991) studies of CheY from *E. coli,* the 3D structure was identified. CheY contains 5 parallel-stranded β-sheets in between five α-helices (Stock et al., 1989). The primary phosphorylation site, aspartate 57, is located at the end of the β3 sheet. Asp57 along with Asp12 and 13 form the "aspartate triad" that is responsible for coordination with Mg^{2+} ion necessary for stabilization of the transitional pentavalent phosphate state

Table 3. CheY proteins, CheY-like domains and response regulators similar to CheY used in analysis.

Gene Name	Organism	GenBank Identification Number	Gene Name	Organism	GenBank Identification Number
CheY	*Escherichia coli*	145525	CheB	*Agrobacterium tumefaciens*	3282794
CheY	*Salmonella typhimurium*	116292	CheB	*Rhizobium meliloti*	534840
CheY	*Burkholderia pseudomallei*	1688294	CheB	*Escherichia coli*	145524
CheY	*Pseudomonas putida*	2853598	CheB	*Salmonella typhimurium*	116280
CheY	*Pseudomonas aeruginosa*	2500739	CheB	*Pseudomonas aeruginosa*	3241970
CheY	*Helicobacter pylori*	4103138	CheB1	*Borrelia burgdorferi*	2688322
CheY	*Campylobacter jejuni*	3413457	CheB2	*Borrelia burgdorferi*	2688489
difD	*Myxococcus xanthus*	3342526	frzZ	*Myxococcus xanthus*	2947294
CheY2	*Rhodobacter sphaeroides*	534996	cheAY	*Rodospirillum centenum*	1621285
CheY2	*Caulobacter crescentus*	3387367	cheA	*Helicobacter pylori*	2313493
CheY3	*Caulobacter crescentus*	3387370	frzE	*Myxococcus xanthus*	120546
CheY1	*Agrobacterium tumefaciens*	3282795	cheV2	*Helicobacter pylori*	4103146
CheY2	*Rhizobium meliloti*	534841	cheV	*Bacillus subtilis*	584926
CheY1	*Borrelia burgdorferi*	2688460	cheV1	*Helicobacter pylori*	2313092
CheY	*Rodospirillum centenum*	1621287	cheA	*Synechocystis*	1001298
CheY	*Halobacterium salinarium*	994802	1EHC	*Escherichia coli*	2194080
CheY	*Archaeglobus fulgidus*	2649557	CCDB	*Bacillus subtilis*	1176639
CheY	*Pyrococcus horikoshii*	3256887	AF1898	*Archaeglobus fulgidus*	2648641
CheY	*Listeria monocytogenes*	620085	AF2249	*Archaeglobus fulgidus*	2648277
CheY	*Thermotoga maritima*	940149	AF1384	*Archaeglobus fulgidus*	2649189
CheY	*Bacillus subtilis*	116278	AF0449	*Archaeglobus fulgidus*	2650175
CheY3	*Borrelia burgdorferi*	2688604	DEGU	*Bacillus subtilis*	118438
CheY	*Treponema denticola*	3493638	lemA	*Pseudomonas aeruginosa*	2623815
CheY	*Treponema pallidum*	1765976	lemA	*Pseudomonas syringae*	1346440
CheY2	*Borrelia burgdorferi*	2688488	rtpA	*Pseudomonas tolaasii*	3953516
CheY4	*Rhodobacter sphaeroides*	personal[a]	Lyase	*Pseudomonas viridiflava*	463195
CheY1	*Caulobacter crescentus*	3387362	divK	*Caulobacter crescentus*	2120415
CheY2	*Agrobacterium tumefaciens*	3282791	rprY	*Bacteroides fragilis*	485521
CheY1	*Rhizobium meliloti*	534836	resD	*Bacillus subtilis*	466194
CheY	*Synechocystis*	1001302	phoP	*Bacillus subtilis*	2293270

Table 3 (continued)

CheY	*Synechocystis*	1001303	rpr2	*Lactobacillus sakei*	4104599
CheB	*Rodospirillum centenum*	1621288	phoB	*Escherichia coli*	130119
CheB	*Rhodobacter sphaeroides*	2293004	phoB	*Shigella dysenteriae*	1172482
CheB	*Treponema pallidum*	3322930	phoB	*Shigella flexneri*	1172483
CheB	*Halobacterium salinarium*	994803	phoB	*Klebsiella pneumoniae*	1172481
CheB	*Archaeglobus fulgidus*	2649549	phoB	*Vibrio cholerae*	3282774
CheB	*Pseudomonas aeruginosa*	3721571	phoB	*Pseudomonas aeruginosa*	130120
CheB	*Pseudomonas putida*	2853601	phoB	*Haemophilus influenzae*	1172480
CheB	*Bacillus subtilis*	2634014	YC27	*Pseudomonas aeruginosa*	129159
CheB	*Pyrococcus horikoshii*	3256888	AFQ1	*Streptomyces coelicolor*	543777
CheB	*Caulobacter crescentus*	3387366	RV0981	*Mycobacterium tuberculosis*	2916942

[a] - J. P. Armitage, Personal communication

(Bellsolell et al., 1994). Two other residues make up the active site: threonine 87, which acts as a general acid-base catalyst and lysine 109, which forms a salt bridge with Asp57 (Moy et al., 1994) (Figure 7).

All of the retrieved sequences have the same general fold. Gaps introduced into the alignment (Figure 8) fall within loops connecting the secondary structure elements as previously observed in other response regulators (Morel-Deville et al., 1998). The N-terminal region appears to contain folding conservation while only a few of the residues are actually identical. An aspartate residue corresponding to the primary phosphorylation site in *E. coli* (Asp57) is well conserved (Figure 8). The other two aspartate residues in the "aspartate triad" (Asp 12 and Asp13) are similarly conserved.

C. Identification of critical residues for CheA, CheZ and FliM binding

CheA binding sites overall appear to be conserved; but some, like Ala99, are only conserved in CheY and the CheB domains, not the other domains or response regulators. Thus, this residue may serve as a "signature" residue for interaction with CheA. The CheY-like N-terminal portion is where the CheB protein gets the phosphoryl group from CheA (Djordjevic et al., 1998). Similarly, other studies have found that the CheY domain of CheB contains side chains similar to those observed in CheY (Stewart, 1993; Stock and Surette, 1996). Since these findings suggest that the activation-induced conformational changes in CheB are similar to those in CheY, one would expect a conservation of the critical residues implicated in phosphorylation activity. The fact that some residues are not conserved in the other response regulators and a few CheY domains suggest either a different mode of binding/activation or that only certain

Figure 7. Mapping of critical residues to CheY crystallized structure. Top view of the crystallized CheY protein shows the five 5 parallel-stranded β-sheets that comprise the active site surrounded by five α helices. D57, the primary phosphorylation site, together with D12 and D13 form the "aspartate triad" that interacts with the Mg^{2+} ion. T87 acts as a general acid-base catalyst and L109 forms the salt bridge with D57. The C-terminus starts beneath the bottom right α-helix. Crystallized structure is based on NMR (Moy et al., 1994) and X-Ray crystallography (Volz and Matsumura, 1991) studies. Model was generated using the Swiss-Pdb viewer (http://www.expasy.ch/spdbv/mainpage.html). Red, α-helix; yellow, β-sheets; gray, loops; D, aspartate; T, threonine; L, leucine.

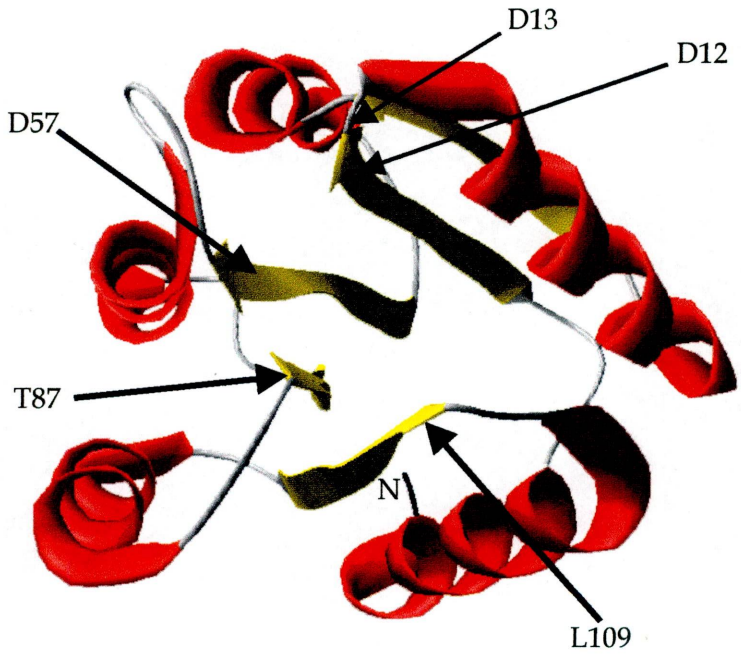
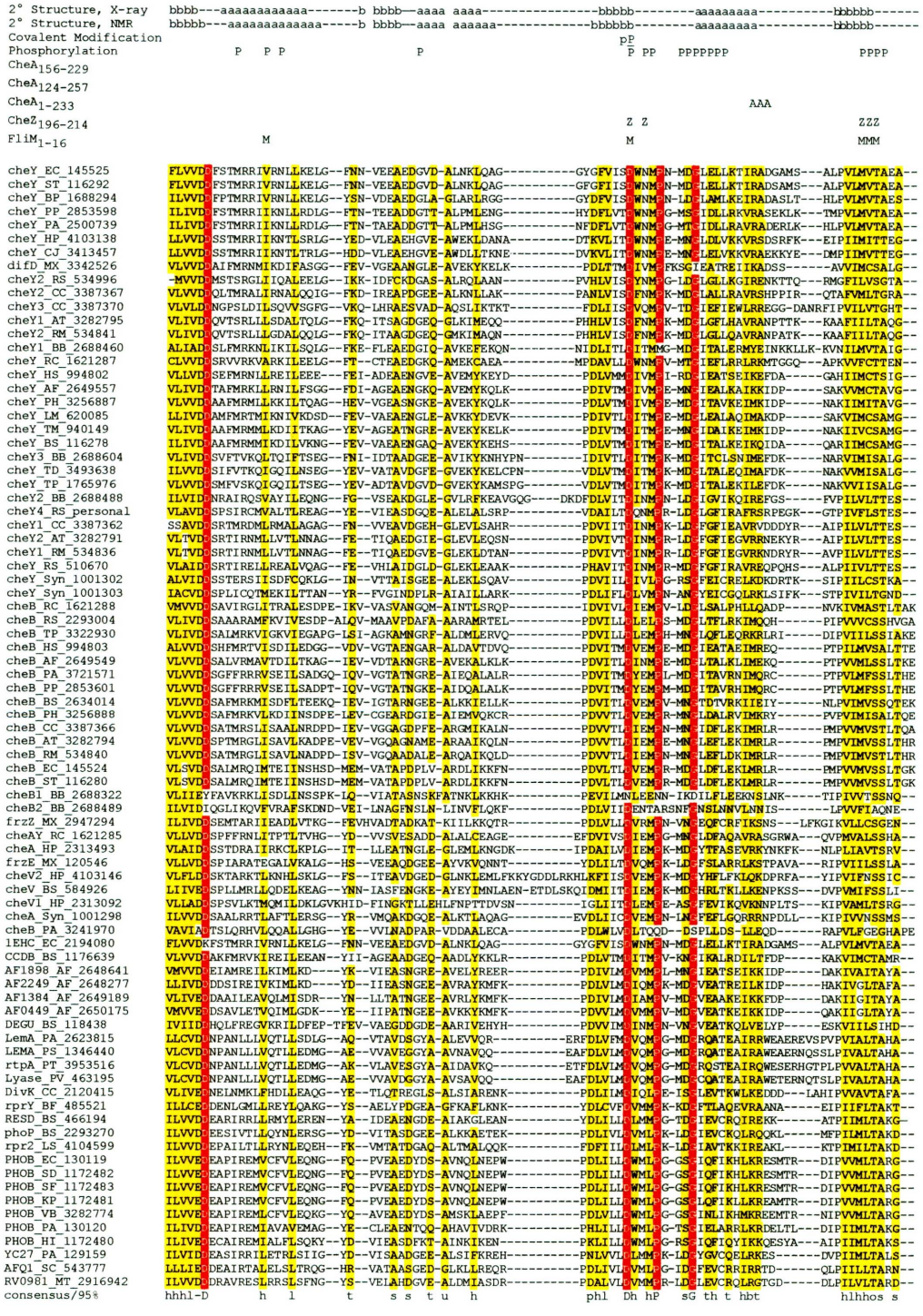
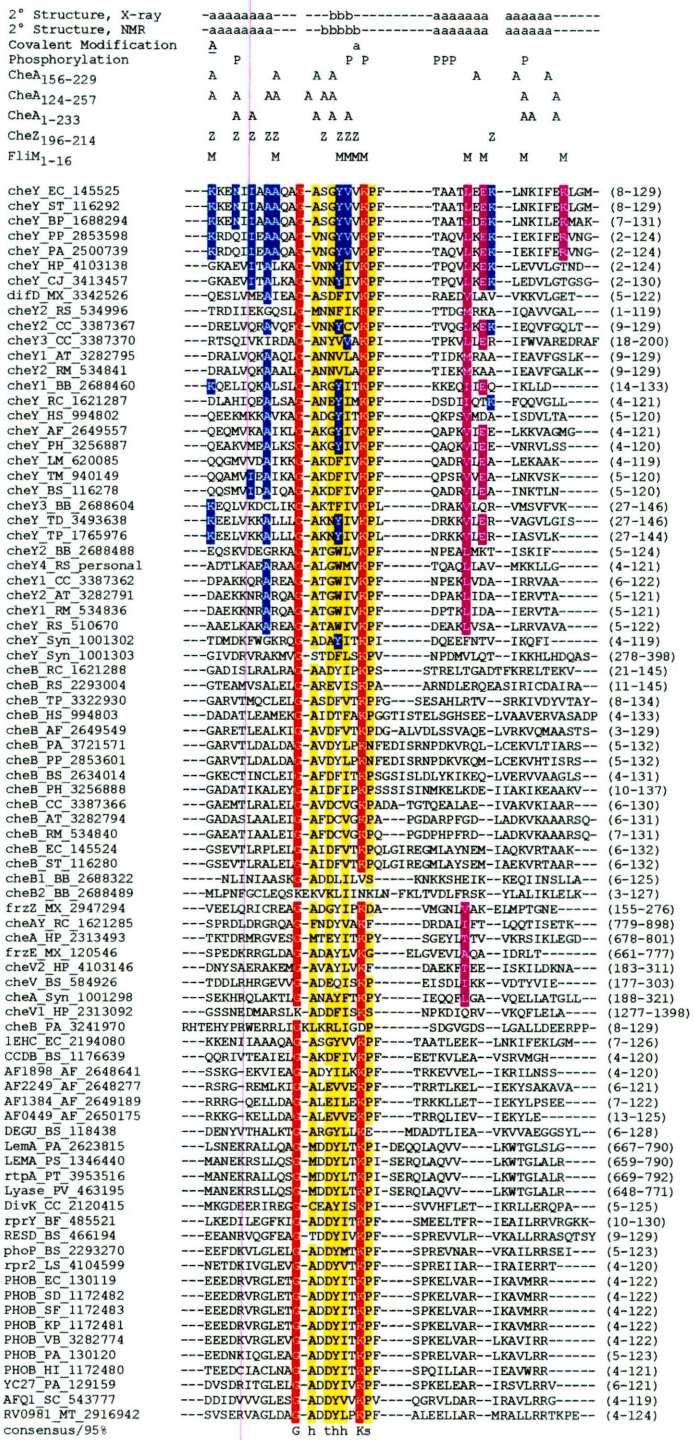

D13
D12
D57
T87
N
L109

Figure 8. Multiple alignment of CheY proteins, CheY-like domains and response regulators similar to CheY. Sequences were retrieved through BLAST (Altschul et al., 1997) searches of non-redundant databases using all known CheY sequences as bait, using E-value as a cut-off parameter. Sequences were aligned using Clustal X (Jeanmougin et al., 1998; Thompson et al., 1997) and manual alignment. Similar amino acid residues are highlighted in yellow; identical residues are highlighted in red. Residues were colored according to a 95% consensus (generated at http://www.bork.embl-heidelberg.de/Alignment/consensus.html [N. Brown and J. Lai, unpublished]); −, negatively charged residues (D and E); b, big residues (E, F, I, K, L, M, Q, R, W and Y); h, hydrophobic residues (A, C, F, I, L, M, V, W and Y); l, aliphatic residues (I, L and V); p, polar residues (C, D, E, H, K, N, Q, R, S and T); s, small residues (A, C, D, G, N, P, S, T and V); t, turnlike residues (A, C, D, E, G, H, K, N, Q, R, S and T; u, tiny residues (A, G and S). Residues important for CheZ binding are highlighted in blue; those important for FliM binding are highlighted in pink. The left column indicates protein/gene name, species abbreviation and GenBank protein identification numbers separated by underscores. Secondary structure elements are based on NMR (Moy et al., 1994) and X-ray crystallography (Volz and Matsumura, 1991) studies of *E. coli* CheY: a, α-helix; b, β-sheet; -, loop. Annotation of residues that undergo covalent modification (Ramakrishnan et al., 1998; Sanders et al., 1989; Appleby and Bourret, 1999; Stock et al., 1985) and those important for CheA (McEvoy et al., 1998; Welch et al., 1998; Swanson et al., 1995), CheZ and FliM (McEvoy et al., 1999) binding are also based on *E. coli* CheY studies; P, 1° phosphorylation site; p, 2° phosphorylation site; A, 1° acetylation site; a, 2° acetylation site; P, residue involved in phosphorylation; A, CheA binding residue; Z, CheZ binding residue; M, FliM binding residue. See Table 3 for definition of species abbreviations.


2° Structure, X-ray
2° Structure, NMR
Covalent Modification
Phosphorylation
CheA156-229
CheA124-257
CheA1-233
CheZ196-214
FliM1-16
cheY_EC_145525
cheY_ST_116292
cheY_BP_1688294
cheY_PP_2853598
cheY_PA_2500739
cheY_HP_4103138
cheY_CJ_3413457
difD_MX_3342526
cheY2_RS_534996
cheY2_CC_3387367
cheY3_CC_3387370
cheY1_AT_3282795
cheY2_RM_534841
cheY1_BB_2688460
cheY_RC_1621287
cheY_HS_994802
cheY_AF_2649557
cheY_PH_3256887
cheY_LM_620085
cheY_TM_940149
cheY_BS_116278
cheY3_BB_2688604
cheY_TD_3493638
cheY_TP_1765976
cheY2_BB_2688488
cheY4_RS_personal
cheY1_CC_3387362
cheY2_AT_3282791
cheY1_RM_534836
cheY_RS_510670
cheY_Syn_1001302
cheY_Syn_1001303
cheB_RC_1621288
cheB_RS_2293004
cheB_TP_3322930
cheB_HS_994803
cheB_AF_2649549
cheB_PA_3721571
cheB_PP_2853601
cheB_BS_2634014
cheB_PH_3256888
cheB_CC_3387366
cheB_AT_3282794
cheB_RM_534840
cheB_EC_145524
cheB_ST_116280
cheB1_BB_2688322
cheB2_BB_2688489
frzZ_MX_2947294
cheAY_RC_1621285
cheA_HP_2313493
frzE_MX_120546
cheV2_HP_4103146
cheV_BS_584926
cheV1_HP_2313092
cheA_Syn_1001298
cheB_PA_3241970
1EHC_EC_2194080
CCDB_BS_1176639
AF1898_AF_2648641
AF2249_AF_2648277
AF1384_AF_2649189
AF0449_AF_2650175
DEGU_BS_118438
LemA_PA_2623815
LEMA_PS_1346440
rtpA_PT_3953516
Lyase_PV_463195
DivK_CC_2120415
rprY_BF_485521
RESD_BS_466194
phoP_BS_2293270
rpr2_LS_4104599
PHOB_EC_130119
PHOB_SD_1172482
PHOB_SF_1172483
PHOB_KP_1172481
PHOB_VB_3282774
PHOB_PA_130120
PHOB_HI_1172480
YC27_PA_129159
AFQ1_SC_543777
RV0981_MT_2916942
consensus/95%

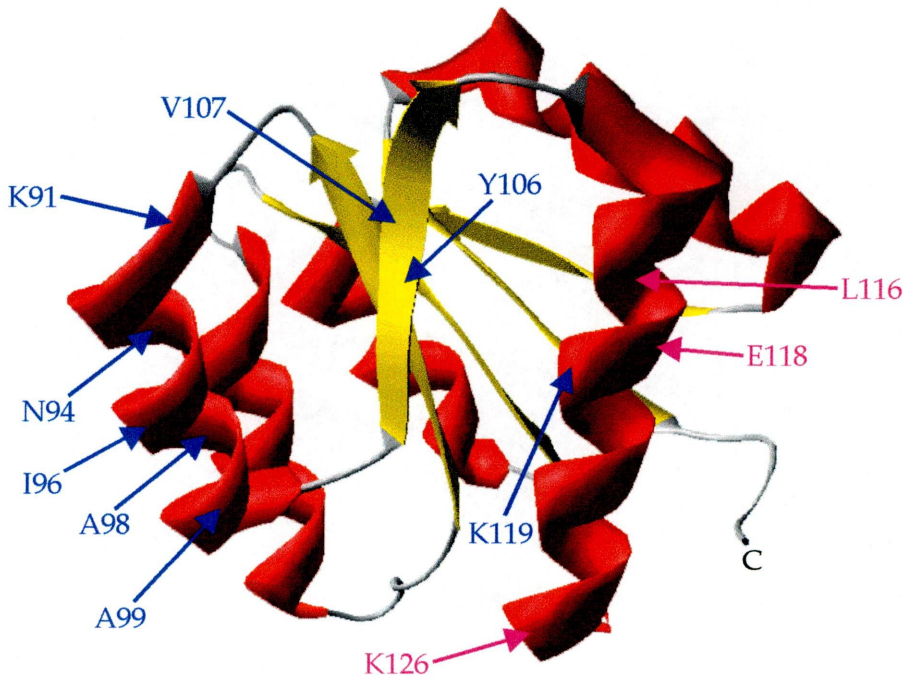

2° Structure, X-ray -aaaaaaaa--- ---bbb---------aaaaaaa aaaaaa------
2° Structure, NMR -aaaaaaaa--- --bbbbb---------aaaaaaa aaaaaa------
Covalent Modification A a
Phosphorylation P P P PPP P
CheA156-229 A A A A A A A
CheA124-257 A A AA A AA A A
CheA1-233 A A A A A A A
CheZ196-214 Z Z Z ZZ Z ZZZ Z
FliM1-16 M M MMMM M M M M
cheY_EC_145525 ---KKENIIAAAQAE-ASGYVVKPF------TAATLEEK--LNKIFEKLGM- (8-129)
cheY_ST_116292 ---KKENIIAAAQAE-ASGYVVKPF------TAATLEEK--LNKIFEKLGM- (8-129)
cheY_BP_1688294 ---KKENIIAAAQAE-ASGYVVKPF------TAATLDEK--LNKILEKMAK- (7-131)
cheY_PP_2853598 ---KRDQIIAAAQAE-VNGYVVKPF------TAQVLKER--IEKIFEKVNG- (2-124)
cheY_PA_2500739 ---KRDQIIAAAQAE-VNGYVVKPF------TAQVLKER--IEKIFEKVNG- (2-124)
cheY_HP_4103138 ---GKAEVIIALKAE-VNNYIVKPF------TPQVLKER--LEVVLGTND-- (2-124)
cheY_CJ_3413457 ---GKAEVIIALKAE-VNNYIVKPF------TPQVLKER--LEDVLGTGSG- (2-130)
difD_MX_3342526 ---QESLVMEAIEAE-ASDFIVKPF------RAEDVLAV--VKKVLGET--- (5-122)
cheY2_RS_534996 ---TRDIIEKGQSLE-MNNFIKKPF------TTDGLRKA--IQAVVGAL--- (1-119)
cheY2_CC_3387367 ---DRELVQRAVQFE-VNNYCVKPF------TVQGLKER--IEQVFGQLT-- (9-129)
cheY3_CC_3387370 ---RTSQIVKIRDAE-ANYVVAKPI------TPKVLQR--IFWVAREDRAF (18-200)
cheY1_AT_3282795 ---DRALVQKAAQLE-ANNVLAKPF------TIDKVRAA--IEAVFGSLK-- (9-129)
cheY2_RM_534841 ---DRALVQKAALLE-ANNVLAKPF------TIEKVKAA--IEAVFGALK-- (9-129)
cheY1_BB_2688460 ---KQELIQKALSLE-ARGYITKPF------KKEQILQQ--IKLLD------ (14-133)
cheY_RC_1621287 ---DLAHIQKALSAE-ANKYIMKPF------DSDIIQTK--FQQVGLL---- (4-121)
cheY_HS_994802 ---QEEKMKKAVKAE-ADGYITKPF------QKPSVMDA--ISDVLTA---- (5-120)
cheY_AF_2649557 ---QEQMVKAAIKLE-AKGYIVKPF------QAPKVLEE--LKKVAGMG--- (4-121)
cheY_PH_3256887 ---QEAKVMEAIKSE-AKGYIVKPF------QAQKVLEE--VNRVLSS---- (4-120)
cheY_LM_620085 ---QQGMVVDAIKKE-AKDFIVKPF------QADRVLEA--LEKAAK----- (4-119)
cheY_TM_940149 ---QQAMVIEAIKAE-AKDFIVKPF------QPSRVVEA--LNKVSK----- (5-120)
cheY_BS_116278 ---QQSMVIEAIQAE-AKDFIVKPF------QADRVLEA--INKTLN----- (5-120)
cheY3_BB_2688604 ---KEQLVKDCLIKE-AKTFIVKPL------DRAKVLQR--VMSVFVK---- (27-146)
cheY_TD_3493638 ---KEELVKKALLLE-AKNYIVKPL------DRKKVLER--VAGVLGIS--- (27-146)
cheY_TP_1765976 ---KEELVKKALLLE-AKNYIVKPL------DRKKVLER--IASVLK----- (27-144)
cheY2_BB_2688488 ---EQSKVDEGRKAE-ATGWLVKPF------NPEAIMKT--ISKIF------ (5-124)
cheY4_RS_personal ---ADTLKAERAAAE-ALGWMVKPF------TQAQVLAV--MKKLLG----- (4-121)
cheY1_CC_3387362 ---DPAKKQRAREAE-ATGWIVKPF------NPEKVVDA--IRRVAA----- (6-122)
cheY2_AT_3282791 ---DAEKKNRARQAE-ATGWIVKPF------DPAKVIDA--IERVTA----- (5-121)
cheY1_RM_534836 ---DAEKKNRARQAE-ATGWIVKPF------DPTKVIDA--IERVTA----- (5-121)
cheY_RS_510670 ---AAELKAKAREAE-ATAWIVKPF------DEAKEVSA--LRRVAVA---- (5-122)
cheY_Syn_1001302 ---TDMDKFWGKRQE-ADAYITKPI------DQEEFNTV--IKQFI------ (4-119)
cheY_Syn_1001303 ---GIVDRVRAKMVE-STDFLSKPV------NPDMVLQT--IKKHLHDQAS- (278-398)
cheB_RC_1621288 ---GADISLRALRAE-AADYIPKPS-----STRELTGADTFKRELTEKV--- (21-145)
cheB_RS_2293004 ---GTEAMVSALELE-AREVISKPA-----ARNDLERQEASIRICDAIRA-- (11-145)
cheB_TP_3322930 ---GARVTMQCLELE-ASDFVTKPFG---SESAHLRTV--SRKIVDYVTAY- (8-134)
cheB_HS_994803 ---DADATLEAMEK--AIDTFAKPGGTISTELSGHSEE-LVAAVERVASADP (4-133)
cheB_AF_2649549 ---GARETLEALKE--AVDFVTKPDG-ALVDLSSVAQE-LVRKVQMAASTS- (3-129)
cheB_PA_3721571 ---GARVTLDALDA--AVDYLPKNFEDISRNPDKVRQL-LCEKVLTIARS-- (5-132)
cheB_PP_2853601 ---GARVTLDALDA--AVDYLPKNFEDISRNPDKVKQM-LCEKVHTISRS-- (5-132)
cheB_BS_2634014 ---GKECTINCLEI--AFDFITKPSGSISLDLYKIKEQ-LVERVVAAGLS-- (4-131)
cheB_PH_3256888 ---GADATIKALEY--AIDFIPKPSSSISINMKELKDE-IIAKIKEAAKV-- (10-137)
cheB_CC_3387366 ---GAEMTLRALELE-AVDCVGKPADA-TGTQEALAE--IVAKVKIAAR--- (6-130)
cheB_AT_3282794 ---GADASLAALEIE-AFDCVGKPA---PGDARPFGD--LADKVKAAARSQ- (6-131)
cheB_RM_534840 ---GAEATIAALEIE-AFDCVGKPQ---PGDPHPFRD--LADKVKAAARSQ- (7-131)
cheB_EC_145524 ---GSEVTLRPLELE-AIDFVTKPQLGIREGMLAYNEM-IAQKVRTAAK--- (6-132)
cheB_ST_116280 ---GSEVTLRALELE-AIDFVTKPQLGIREGMLAYSEM-IAEKVRTAAR--- (6-132)
cheB1_BB_2688322 -----NLINIAASKE-ADDLILVS-----KNKKSHEIK--KEQIINSLLA-- (6-125)
cheB2_BB_2688489 ---MLPNFGCLEQSKEKVKLIINKLN-FKLTVDLFRSK--YLALIKLELK-- (3-127)
frzZ_MX_2947294 ---VEELQRICREAE-ADGYIPKDA-----VMGNLIAK-ELMPTGNE----- (155-276)
cheAY_RC_1621285 ---SPRDLDRGRQAE-FNDYVAKPF------DRDAILFT--LQQTISETK--- (779-898)
cheA_HP_2313493 ---TKTDRMRGVESE-MTEYITKPY-----SGEYLETV--VKRSIKLEGD-- (678-801)
frzE_MX_120546 ---SPEDKRRGLDAE-ADAYLVKG-----ELGVEVLAQA--IDRLT------- (661-777)
cheV2_HP_4103146 ---DNYSAERAKEME-AVAYLVKF------DAEKFIEE--ISKILDKNA--- (183-311)
cheV_BS_584926 ---TDDLRHRGEVVE-ADEQISKP------EISDLIKK--VDTYVIE----- (177-303)
cheA_Syn_1001298 ---SEKHRQLAKTLE-ANAYFTKPY-----IEQQFVGA--VQELLATGLL-- (188-321)
cheV1_HP_2313092 ---GSSNEDMARSLK-ADDFISKS------NPKDIQRV--VKQFLELA---- (1277-1398)
cheB_PA_3241970 RHTEHYPRWERRLIKLKRLIGDP-------SDGVGDS--LGALLDEERPP- (8-129)
1EHC_EC_2194080 ---KKENIIAAAQAE-ASGYVVKPF----TAATLEEK--LNKIFEKLGM--- (7-126)
CCDB_BS_1176639 ---QQRIVTEAIELE-AKDFIVKPF----EETKVLEA--VSRVMGH------ (4-120)
AF1898_AF_2648641 ---SSKG-EKVIEAE-ADYILKKPF----TRKEVVEL--IKRILNSS----- (4-120)
AF2249_AF_2648277 ---RSRG-REMLKIE-ALEVVEKPF----TRRTLKEL--IEKYSAKAVA--- (6-121)
AF1384_AF_2649189 ---RRRG-QELLDAE-ALEILEKPF----TRKKLLET--IEKYLPSEE---- (7-122)
AF0449_AF_2650175 ---RKKG-KELLDAE-ALEVVEKPF----TRRQLIEV--IEKYLE------- (13-125)
DEGU_BS_118438 ---DENYVTHALKTE-ARGYLLKE----MDADTLIEA--VKVVAEGGSYL-- (6-128)
LemA_PA_2623815 ---LSNEKRALLQAE-MDDYLTKPI-DEQQLAQVV----LKWTGLSLG---- (667-790)
LEMA_PS_1346440 ---MANEKRSLLQSE-MDDYLTKPI-SERQLAQVV----LKWTGLALR---- (659-790)
rtpA_PT_3953516 ---MANEKRALLQSE-MDDYLTKPI-SERQLAQVV----LKWTGLALR---- (669-792)
Lyase_PV_463195 ---MANEKRSLLQSE-MDDYLTKPI-SERQLAQVV----LKWTGLALR---- (648-771)
DivK_CC_2120415 ---MKGDEERIREG-CEAYISKPI----SVVHFLET--IKRLLERQPA--- (5-125)
rprY_BF_485521 ---LKEDILEGFKIE-ADDYITKPF----SMEELTFR--IEAILRRVRGKK- (10-130)
RESD_BS_466194 ---EEANRVQGFEAE-TDDYIVKPF----SPREVVLR--VKALLRRASQTSY (9-129)
phoP_BS_2293270 ---EEFDKVLGLELE-ADDYMTKPF----SPREVNAR--VKAILRRSEI--- (5-123)
rpr2_LS_4104599 ---NETDKIVGLEVE-ADDYVTKPF----SPREIIAR--IRAIERRT----- (4-120)
PHOB_EC_130119 ---EEEDRVRGLETE-ADDYITKPF----SPKELVAR--IKAVMRR------ (4-122)
PHOB_SD_1172482 ---EEEDRVRGLETE-ADDYITKPF----SPKELVAR--IKAVMRR------ (4-122)
PHOB_SF_1172483 ---EEEDRVRGLETE-ADDYITKPF----SPKELVAR--IKAVMRR------ (4-122)
PHOB_KP_1172481 ---EEEDRVRGLETE-ADDYITKPF----SPKELVAR--IKAVMRR------ (4-122)
PHOB_VB_3282774 ---EEEDKVRGLEVE-ADDYITKPF----SPKELVAR--LKAVIRR------ (4-122)
PHOB_PA_130120 ---EEDNKIQGLEAE-ADDYITKPF----SPRELVAR--LKAVLRR------ (5-123)
PHOB_HI_1172480 ---TEEDCIACLNAE-ADDYITKPF----SPQILLAR--IEAVWRR------ (4-121)
YC27_PA_129159 ---DVSDRITGLEIE-ADDYIVKPF----SPKELEAR--IRSVLRRV----- (6-121)
AFQ1_SC_543777 ---DDIDVVVGLESE-ADDYVVKPV----QGRVLDAR--IRAVLRRG----- (4-119)
RV0981_MT_2916942 ---SVSERVAGLDAE-ADDYLPKPF----ALEELIAR--MRALLRRTKPE-- (4-124)
consensus/95% G h thh Ks

residues are critical for CheA binding/activation. The high degree of conservation in the N-terminal indicates a common mechanism for phosphorylation and associated conformational changes.

While the N-terminal of CheY is primarily involved in phosphorylation, the C-terminal is where the main interaction with CheA, CheZ or FliM takes place. Residues that interact with CheZ appear to be highly conserved in γ-bacteria but not in other bacterial classes nor those that lack a CheZ phosphatase. Based on the alignment we were able to derive the consensus sequence K-(K/R)-(E/D)-(N/Q)-I-I-A-A-A-Q-A-G-(A/V)-(S/N)-G-Y-V starting from Lys91 in *E. coli*. Underlined residues represent key residues in CheZ-CheY binding interactions. The consensus residues fall on the α4 helix and the β5 sheet (Figure 9), near Lys109, which forms the salt-bridge with Asp57. Binding of CheZ to this region may disrupt this bridge, promoting the removal of the phosphoryl group.

CheY interacts with the FliM flagellar motor protein. Critical residues needed for this interaction were identified in NMR studies (Moy et al., 1994), and we mapped these residues onto the CheY alignment (Figure 8). Some, like Glu118, appear to be conserved in all CheY proteins but not in CheY-like domains or other response regulators. This suggests that CheY-like domains or other response regulators may not interact with the flagella. Comparison of different FliM proteins from many species shows that there are highly conserved regions (i.e. the N-terminus) and highly variable regions (Mathews et al., 1998); and some sequences themselves can be quite different, like FliM from *Agrobacterium tumefaciens* (Deakin et al., 1997). Previous domain analysis of FliM has

Figure 9. Mapping of predicted critical CheZ and FliM binding residues. Side view of the crystallized CheY protein showing the slight pocket formed by the β5 sheet in between the α4 and α5 helices. Blue residues represent those predicted to be important for CheZ binding. Pink residues are important for FliM binding. The N-terminus is located after the α5 helix. Crystallized structure is based on NMR (Moy et al., 1994) and X-Ray crystallography (Volz and Matsumura, 1991) studies. Model was generated using the Swiss-Pdb viewer (http://www.expasy.ch/spdbv/mainpage.html). Red, α-helix; yellow, β-sheets; gray, loops; A, alanine; E, glutamate; I, isoleucine; K, lysine; L, leucine; N, asparagine; V, valine; Y, tyrosine.

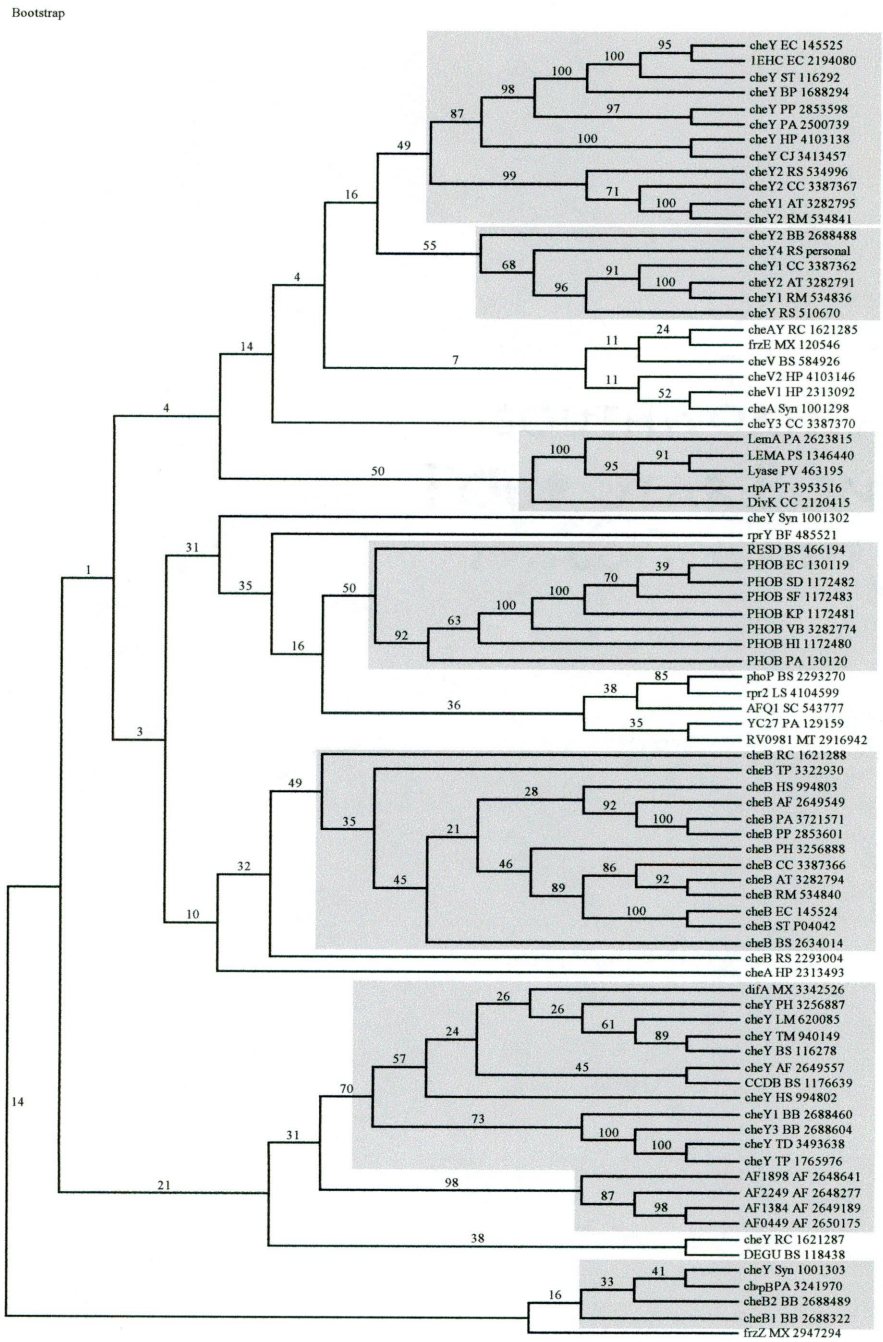

V107
K91
Y106
L116
E118
N94
I96
A98
A99
K119
C
K126

revealed the importance of the N-terminal for CheY and CheY~P binding (Bren and Eisenbach, 1998; Toker and Macnab, 1997). More recent studies have suggested that regions in the C-terminal and possibly the middle of FliM are also important for CheY and CheY~P binding (Mathews et al., 1998). CheY-like domains and other response regulators may interact with FliM in poorly conserved regions, making it difficult to identify FliM-binding proteins.

Phylogenetic analysis of known and putative CheY and CheY domains was also performed. Phylogenetic trees were created utilizing the PAUP software package (Swofford, 1991), PHYLIP (Felsenstein, 1989) and Clustal X (Jeanmougin et al., 1998). As shown in Figure 10 the main (FliM-interacting) CheYs are clustered together in one branch, while the multiple copy CheYs are grouped together in a separate branch. This supports the notion that the multiple copy CheY proteins have different functions, like acting as phosphate sinks. The CheY-like domains cluster together as well suggesting similar function for these domains which is different from that of the main CheY proteins.

The CheY domains of CheB1 (accession number 2688322) and CheB2 (accession number 2688489) from *Borrelia burgdorferi*, while similar overall, appear to be missing most of the highly conserved residues that comprise the active site. CheB1 is missing many of the necessary residues that are important for proper function (i.e. Asp57 and Lys109) and is lacking conservation of other residues that help make up the active site (i.e. Asp12 and Asp13). Without an active site, CheB1 can not be phosphorylated in the normal way and therefore may not be functional. CheB2 is also missing certain critical

Figure 10. Phylogenetic tree of CheY proteins, CheY-like domains and similar response regulators. The tree was generated (by Jonathan Eisen, collaborator) using the PAUP program (Swofford, 1991). Bootstrap values, indicating the number of times a particular node was found in trees generated from 100 bootstrap replicates, are shown on the tree. Gray boxes represent areas of significance. Protein/gene name, species abbreviation and GenBank protein identification numbers are shown. See Table 3 for definition of species abbreviations.


Bootstrap
cheY EC 145525
1EHC EC 2194080
cheY ST 116292
cheY BP 1688294
cheY PP 2853598
cheY PA 2500739
cheY HP 4103138
cheY CJ 3413457
cheY2 RS 534996
cheY2 CC 3387367
cheY1 AT 3282795
cheY2 RM 534841
cheY2 BB 2688488
cheY4 RS personal
cheY1 CC 3387362
cheY2 AT 3282791
cheY1 RM 534836
cheY RS 510670
cheAY RC 1621285
frzE MX 120546
cheV BS 584926
cheV2 HP 4103146
cheV1 HP 2313092
cheA Syn 1001298
cheY3 CC 3387370
LemA PA 2623815
LEMA PS 1346440
Lyase PV 463195
rtpA PT 3953516
DivK CC 2120415
cheY Syn 1001302
rprY BF 485521
RESD BS 466194
PHOB EC 130119
PHOB SD 1172482
PHOB SF 1172483
PHOB KP 1172481
PHOB VB 3282774
PHOB HI 1172480
PHOB PA 130120
phoP BS 2293270
rpr2 LS 4104599
AFQ1 SC 543777
YC27 PA 129159
RV0981 MT 2916942
cheB RC 1621288
cheB TP 3322930
cheB HS 994803
cheB AF 2649549
cheB PA 3721571
cheB PP 2853601
cheB PH 3256888
cheB CC 3387366
cheB AT 3282794
cheB RM 534840
cheB EC 145524
cheB ST P04042
cheB BS 2634014
cheB RS 2293004
cheA HP 2313493
difA MX 3342526
cheY PH 3256887
cheY LM 620085
cheY TM 940149
cheY BS 116278
cheY AF 2649557
CCDB BS 1176639
cheY HS 994802
cheY1 BB 2688460
cheY3 BB 2688604
cheY TD 3493638
cheY TP 1765976
AF1898 AF 2648641
AF2249 AF 2648277
AF1384 AF 2649189
AF0449 AF 2650175
cheY RC 1621287
DEGU BS 118438
cheY Syn 1001303
chrpBPA 3241970
cheB2 BB 2688489
cheB1 BB 2688322
frzZ MX 2947294

residues for the active site; however it does contain the conserved primary phosphorylation site (Asp57).

Interestingly, *B. burgdorferi* is the only organism that contains multiple copies of CheB, one functional and one non-functional. Since both of these proteins were identified as CheB through complete genome sequence annotation (Fraser et al., 1997) their function hasn't been proven experimentally. CheB1 is found in the operon, which appears to be transferred into *Borrelia* horizontally (Zhulin, 1999), whereas CheB2 is in another operon, which is homologous to that in another spirochaete, *Treponema pallidum*. Thus, it seems that the horizontally transferred operon may not be functional.

In summary, we have performed database searches identifying more than 50 known and putative CheY proteins and CheY-like domains. Alignment analysis revealed an overall conservation in protein folding. Based on published reports, critical residues were identified that played important roles in phosphorylation and interaction with CheA, CheZ and FliM proteins. A consensus sequence predicted to be involved in CheZ binding was derived based on the multiple alignment and analysis of conservation within γ-proteobacteria. Based on the alignment and identification of critical residues, functional predictions for multiple copy CheY and CheY-like proteins were made, supported by phylogenetic analysis.

CHAPTER FOUR
ANALYSIS OF THE METHYL-ACCEPTING CHEMOTAXIS PROTEINS

A. Introduction

Bacteria sense levels of external chemical stimuli through various receptor proteins. In the chemotaxis pathway, this allows the bacteria to sense environmental conditions and migrate towards more favorable niches based on flagellar rotational control. Flagellar rotation fluctuates between two states: a counter-clockwise (CCW) and a clockwise (CW) rotation. The CCW rotation causes smooth swimming in a straight path; while CW rotation results in tumbling with no net movement (Berg, 1993). In the absence of chemical stimuli, the flagella randomly fluctuate between CW and CCW rotation resulting in a random walk. However, if the bacterium senses that it is approaching an attractant or moving away from a repellent, it tends to tumble less frequently. This increases the movement towards the favorable condition, thereby adding a bias to the random walk (Berg and Brown, 1972).

The process begins with the binding of a stimulatory ligand to the sensory domain of the receptor in the periplasmic space (Stock and Surette, 1996). This binding causes a conformational change in the receptor, propagating the signal across the membrane to the cytoplasmic signaling domain (Chervitz et al., 1995; Chervitz and Falke, 1996). It is this signaling domain that mediates the transfer of information from the receptor to the cytoplasm. The histidine kinase CheA forms a tertiary complex with the receptor and the docking protein CheW (Ninfa et al., 1991). CheA is unusual in that it lacks the sensing domain normally present in most sensor kinases (Appleby et al.,

1996), instead using the receptor as its sensing domain. CheA undergoes autophosphorylation (Hess et al., 1988) and transfers this phosphoryl group to the response regulator, CheY (Sanders et al., 1989). Phospho-CheY (CheY~P) interacts with the flagellar switch protein FliM, causing a change in rotation from CCW to CW (Welch et al., 1993). The phosphatase CheZ removes the phosphate group from CheY~P decreasing the amount present in the cytoplasm, slowly restoring CCW rotation (Blat and Eisenbach, 1996).

An attractant binding to the receptor inhibits the kinase activity of the CheA/CheW/receptor complex resulting in a decrease of CheY~P, thereby favoring smooth swimming (CCW rotation) (Borkovich et al., 1989; Ninfa et al., 1991). Kinase activity is also regulated via methylation and demethylation of conserved glutamate/glutamine located in the cytoplasmic signaling domain. Such receptors are called methyl-accepting chemotaxis proteins or MCPs (Springer et al., 1979). An analysis of known MCPs was performed in 1996 (Le Moual and Koshland, 1996). At that time only 29 known and putative MCP proteins were compared. Today the number of protein sequences has greatly increased due in part to quicker sequencing techniques and complete genome sequencing. We decided to re-analyze MCPs incorporating all new MCPs that have been identified since Le Moual and Koshland's original paper. The MCP sequences from a variety of different organisms were used as queries to perform BLAST (Altschul et al., 1997) database searches. Ninty-five known and putative MCPs were identified (Table 4).

The generally accepted model of MCP structure and domain organization is

Table 4. List of known and putative MCPs used in analysis.

Species	Gene Name (protein identification number[a])	Species	Gene Name (protein identification number[a])
Agrobacterium tumefaciens	Unknown (1381805)	*Haloarcula vallismortis*	Htr2 (1170416)
	MCPA (3153186)	*Leptospirillum ferrooxidans*	Icrl (2808645)
	MCP (3282789)	*Myxococcus xanthus*	DifA (3342523)
Archaeglobus fulgidus	tlpC-1(2649560)		FrzCD (1169748)
	tlpC-2 (2649548)	*Natronomonas pharaonis*	Htr2 (1170417)
Borrelia burgdorferi	MCP1 (BB0578)[b]	*Pseudomonas aeruginosa*	MCP (2626835)
	MCP2 (BB0596)[b]		PilJ (1172509)
	MCP3 (BB0597)[b]	*Pyrococcus horikoshii*	MCP (PH0443)[b]
	MCP4 (BB0680)[b]		MCP (PH0479)[b]
	MCP5 (BB0681)[b]		MCP (PH0491)[b]
Bacillus subtilis	MCPA (730002)		MCP 9PH1852)[b]
	MCPB (730003)		MCP (PH1994)[b]
	MCPC (1708962)		MCP (PH1970)[b]
	tlpA (730958)	*Pseudomonas putida*	nahY (4235480)
	tlpB (730959)	*Rhodobacter capsulatus*	MCPA (2126470)
	tlpC (730960)		MCPB (2126471)
	YhfV (2226258)		MCP (3128262)
	YvaQ (2635882)		MCP (3128282)
	YfmS (2116757)	*Rhizobium leguminosarum*	MCPA (780656)
	YoaH (2619023)		MCPB (2564665)
Caulobacter crescentus	MCPA (462577)		MCPC (2665910)
Clostridium thermocellum	MCP (729201)		MCPD (1764196)
	MCP (4235392)	*Rhizobium meliloti*	Y4FA (2497833)
Desulfovibrio gigas	MCP (4235392)		Y4SI (2497833)
Desulfovibrio vulgaris	drcA (544146)		YCH1 (2497835)
	drcH (887858)	*Rhodobacter sphaeroides*	MCPA (1075570)
Escherichia coli	tsr (2367379)	*Synechocystis*	MCP (sll1294)[c]
	tar (1788195)		MCPA (slr1044)[c]
	trg (1787690)		MCP1 (sll0041)[c]
	tap (1788194)		MCP2 (1001299)

Table 4 (continued)

	Aer (1703222)	*Salmonella typhimurium*	tar (1770886)
Transposon Tn1721	MCP (78533)		tcp (400235)
Enterobacter aerogenes	tas (148350)	*Treponema denticola*	dmcA (2914132)
	tse (148349)		dmcB (1805311)
Helicobacter pylori	MCP (HP0099)[b]		MCPA (2352917)
	MCP2 (HP0103)[b]	*Treponema pallidum*	MCP1 (TP0040)[b]
	MCP3 (HP0082)[b]		MCP2-1 (TP0488)[b]
	48KDAg (1840146)		MCP2-2 (TP0639)[b]
Halobacterium salinarium	HtA (1654419)		MCP2-3 (TP0640)[b]
	HtB (1654421)	*Vibrio cholerae*	AcfB (100874)
	HtC (1654423)		HLYB (123206)
	HtD (1654425)		TCP1 (1174620)
	HtF (1654427)		
	HtH (2072795)		
	HtI (1621047)		
	HtR (2648028)		
	Htr2 (1527137)		
	Rho2 (1527138)		
	htrVIII (3015619)		
	htrXII (4104487)		
	htrXIII (4104483)		

[a]All protein identification numbers are GenBank designations unless noted.

[b]Proteins starting with BB, HP, PH and TP are accessible via individual genome databases at TIGR (http://www.tigr.org/tdb).

[c]Proteins starting with slr and sll are accessible via the CyanoBase website (http://www.kazusa.or.jp/cyano/cyano.html).

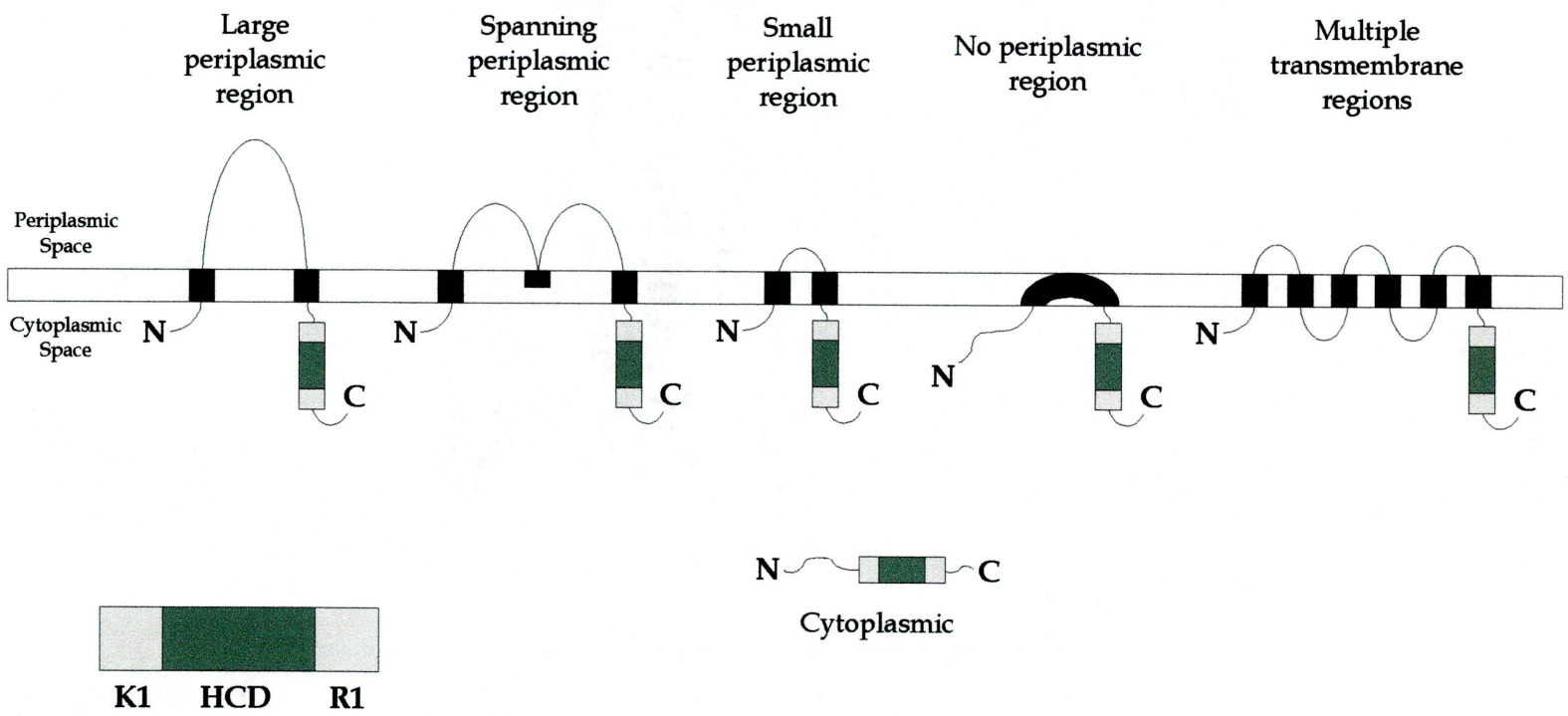
based on NMR and X-ray crystallography studies of the Tar protein from *S. typhimurium* (Milburn et al., 1991). The periplasmic sensory domain is located at the N-terminal and is flanked by two transmembrane regions. The C-terminal contains the signaling domain flanked by two methylation regions (K1 and R1). A linker region connects the sensory and signaling domains. This region contains coiled-coil motifs (Singh et al., 1998) and is believed to help transmit the ligand binding signal from the sensory domain to the signaling domain (Tatsuno et al., 1996; Gardina and Manson, 1996).

However, recent discoveries have identified other MCPs that differ structurally from the generally accepted model. A fifth MCP, Aer, was recently discovered in *E. coli* (Rebbapragada et al., 1997; Bibikov et al., 1997). Aer is unique in that it lacks a periplasmic sensing domain. Instead it senses redox potential via interaction with the electron transport system through a PAS containing sensory domain located in the cytoplasm. Signal transduction studies in *Halobacterium salinarum* have revealed the presence of thirteen different MCPs. These have been classified into three structurally distinct subclasses based on known sequences: 1) Class A - two transmembrane segments, a periplasmic sensing domain and the conserved cytoplasmic signaling domain; B) Class B - two or more transmembrane segments and the conserved cytoplasmic signaling domain and 3) Class C – a soluble protein containing the conserved cytoplasmic signaling domain (Zhang et al., 1996).

B. Topology Prediction

We performed topology studies utilizing the DAS server to classify all identified MCPs. Six different classes were identified based on DAS results (Figure 11): 1) Large

Figure 11. Classification of MCPs according to their predicted membrane topology. All retrieved MCPs can be classified into six different classes based on DAS server predictions (http://www.biokemi.su.se/~server/DAS/). 1) Large periplasmic region, contains a large (~300 amino acid) periplasmic sensing domain; 2) Spanning periplasmic region, the middle part of the periplasmic sensing domain contacts the membrane but does not cross it resulting in an 'M' shape sensing domain; 3) Small periplasmic region, contains a small (~50-80 amino acid) periplasmic sensing domain; 4) No periplasmic region, MCPs are bound to membrane yet the sensing domain is located in the cytoplasm (Aer-like); 5) Multiple transmembrane regions and 6) Cytoplasmic sensors. K1 and R1 represent the two methylation regions and the HCD represents the highly conserved domain.


Large periplasmic region
Spanning periplasmic region
Small periplasmic region
No periplasmic region
Multiple transmembrane regions
Periplasmic Space
Cytoplasmic Space
N
C
N
C
N
C
N
C
N
C
N
C
K1
HCD
R1
N
C
Cytoplasmic

periplasmic region, contains a large (~300 amino acid) periplasmic sensing domain; 2) Spanning periplasmic region, the middle part of the periplasmic sensing domain contacts the membrane but does not cross it resulting in an 'M' shaped sensing domain; 3) Small periplasmic region, contains a small (~50-80 amino acid) periplasmic sensing domain; 4) No-periplasmic region, MCPs are bound to membrane yet the sensing domain is located in the cytoplasm (Aer-like); 5) Multiple transmembrane regions and 6) Cytoplasmic sensor unattached to the membrane.

The largest family are the sensors containing the large periplasmic region and includes well studied MCPs like those from *E. coli* (except Aer), *Enterobacter aerogenes* and *Salmonella typhimurium* (Table 5). This suggests that many MCPs have an overall structure similar to that identified via crystallography studies. The second largest family are the cytoplasmic (presumably, soluble) proteins (Table 5). Since these proteins lack a periplasmic sensing domain, the obvious question is how these transducers sense and transmit sensory signals. A recent report describes the function of a soluble MCP in *H. salinarum*. Named Car, it has been found to play a role in arginine chemotaxis. There are many proposed signaling mechanisms for Car. One involves direct interaction (binding) of arginine to Car in the cytoplasm and relaying of this signal to the chemotaxis pathway. Another states that Car interacts with an unknown cytoplasmic receptor component before any interaction with the chemotaxis pathway (Storch et al., 1999). Interestingly, PSI-BLAST analysis of the two MCPs from *Archaeglobus fulgidus* revealed the presence of a PAS domain (Zhulin and Taylor, 1998). Aer has been found to interact with the electron transport system through its PAS domain, suggesting these

Table 5. Classification of MCPs according to predicted membrane topology.[a]

Species	Protein	Protein Identification Number
1. Large periplasmic region		
Borrelia burgdorferi	MCP2	BB0596
	MCP4	BB0680
	MCP5	BB0681
Bacillus subtilis	McpA	730002
	McpB	730003
	McpC	1708962
	TlpA	730958
	TlpB	730959
	TlpC	730960
	YvaQ	2635882
Caulobacter crescentus	McpA	462577
Desulfovibrio gigas	MCP	4235392
Desulfovibrio vulgaris	DcrA	544146
Enterobacter aerogenes	Tas	148350
	Tse	148349
	McpD	126840
Escherichia coli	Tsr	2367378
	Tar	1788195
	Trg	1787690
	Tap	1788194
Helicobacter pylori	MCP2	HP0103
	MCP3	HP0082
Halobacterium salinarium	HtC	1654423
	HtF	1654427
	Pho2	1527137
Leptospirillum ferrooxidans	Icrl	2808645
Pseudomonas aeruginosa	MCP	2626835
	MCP	2626833
	PctA	1255679
	PilJ	1172509
Pyrococcus horikoshii	MCP	PH0491
Pseudomonas putida	NahY	4235480
Rhodobacter capsulatus	McpA	2126470
	McpB	2126471
Rhizobium leguminosarum	MCP	1684743
	McpA	780656
	McpB	2564665
	McpC	2665910
Rhizobium meliloti	Y4SI	2497834
Salmonella typhimurium	Tcp	400235
	Tar	1170886
Synechocystis	MCP	sll1294
Treponema denticola	DmcA	2914132
	McpA	2352917

Table 5 (continued)

Treponema pallidum	MCP2-1	TP0488
Vibrio cholerae	Acf	431739
	Tcp1	1174620
	HlyB	123206

2. Spanning periplasmic region

Borrelia burgdorferi	MCP3	BB0597
Desulfovibrio vulgaris	DcrH	887858
Helicobacter pylori	MCP	HP0099
Halobacterium salinarium	HtD	1654425
	HtI	1621047
Rhizobium leguminosarum	McpD	1764196
Rhizobium meliloti	Y4FA	2497833

3. Small periplasmic region

Transposon 1721	MCP	78533
Halobacterium salinarium	Htr	2648028
Natronomonas pharaonis	Htr2	1170417
Pyrococcus horikoshii	MCP	PH0443
	MCP	PH1852
Synechocystis	MCP1	sll0041

4. No periplasmic region

Archaeglobus fulgidus	Unknown	1381805
Bacillus subtilis	YoaH	2619023
Clostridium thermocellum	Mcp	729201
Escherichia coli	Aer	1703222
Myxococcus xanthus	DifA	3342523
Pyrococcus horikoshii	Mcp	PH1994
Rhodobacter capsulatus	MCP	3128262
Synechocystis	MCPA	slr1044
Treponema denticola	DmcB	1805311
Treponema pallidum	MCP2	2367665

5. Multiple transmembrane regions

Halobacterium salinarium	HtrVIII	3015619
	Rho2	1527138
Treponema pallidum	MCP1	TP0040
	MCP2-2	TP0639
	MCP2-3	TP0640
	MCP	TP0488

6. Cytoplasmic

Archaeglobus fulgidus	TlpC-1	2649560
	TlpC-2	2649548
Agrobacterium tumefaciens	MCP	3282789
	MCPA	3153186

Table 5 (continued)

Borrelia burgdorferi	MCP1	BB0578
Bacillus subtilis	YhfV	2226258
	YfmS	2116757
Helicobacter pylori	HylB	2313716
	48KDAG	1840146
Halobacterium salinarium	HtA	1654419
	HtB	1654421
	HtH	2072795
	HtrXII	4104487
	HtrXIII	4104483
Haloarcula vallismortis	Htr2	1170416
Myxococcus xanthus	FrzCD	1169748
Pseudomonas aeruginosa	Mcp	3342496
Pyrococcus horikoshii	MCP	PH1970
	MCP	PH0479
Rhizobium meliloti	YCH1	2497835
Rhodobacter sphaeroides	McpA	1075570
Synechocystis	MCP-2	sll0042

[a]All protein identification numbers are GenBank except for those starting with BB, HP, PH and TP (TIGR website) or slr and sll (CyanoBase website).

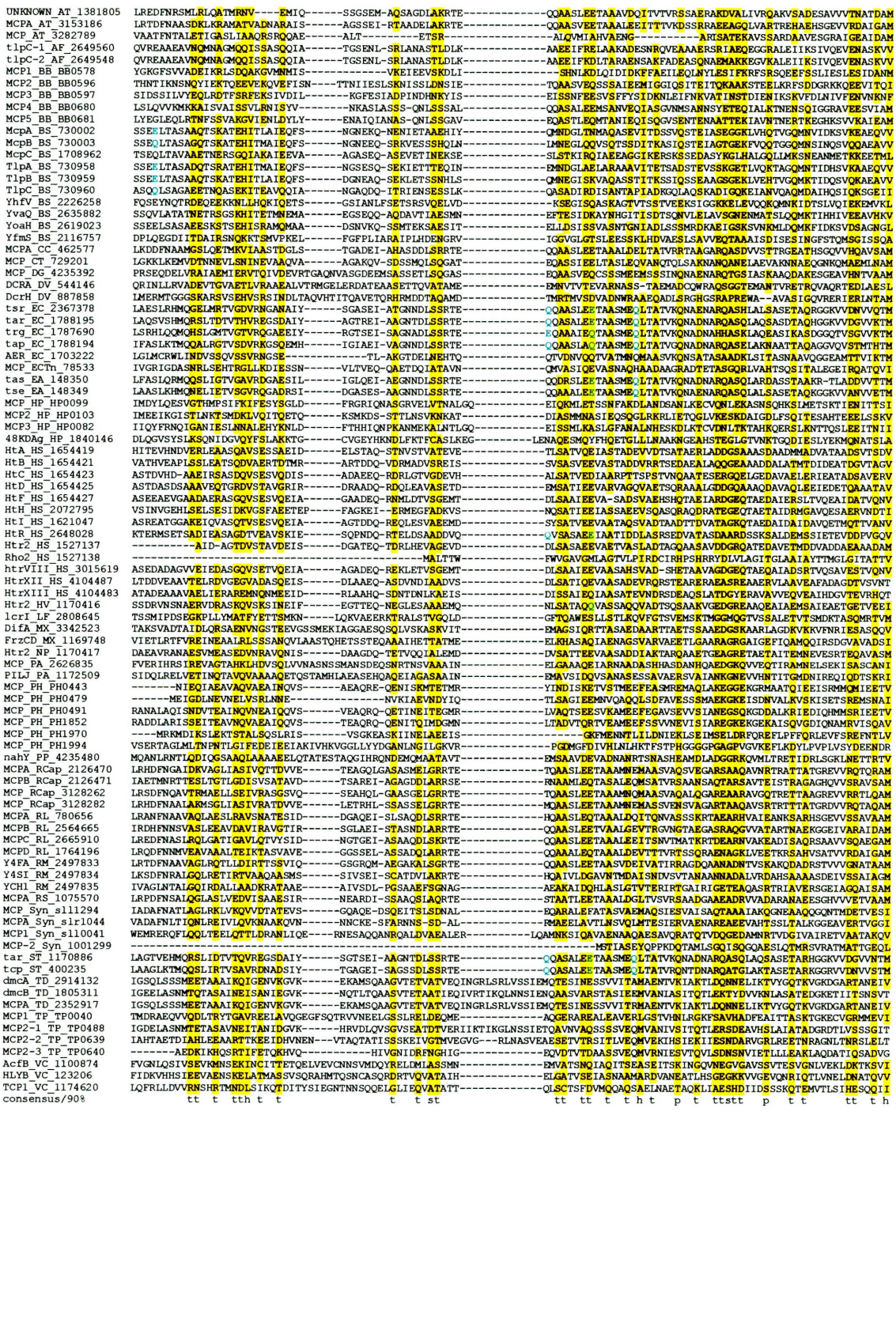
two soluble proteins may associate with the electron transport system in the membrane and monitor redox potential.

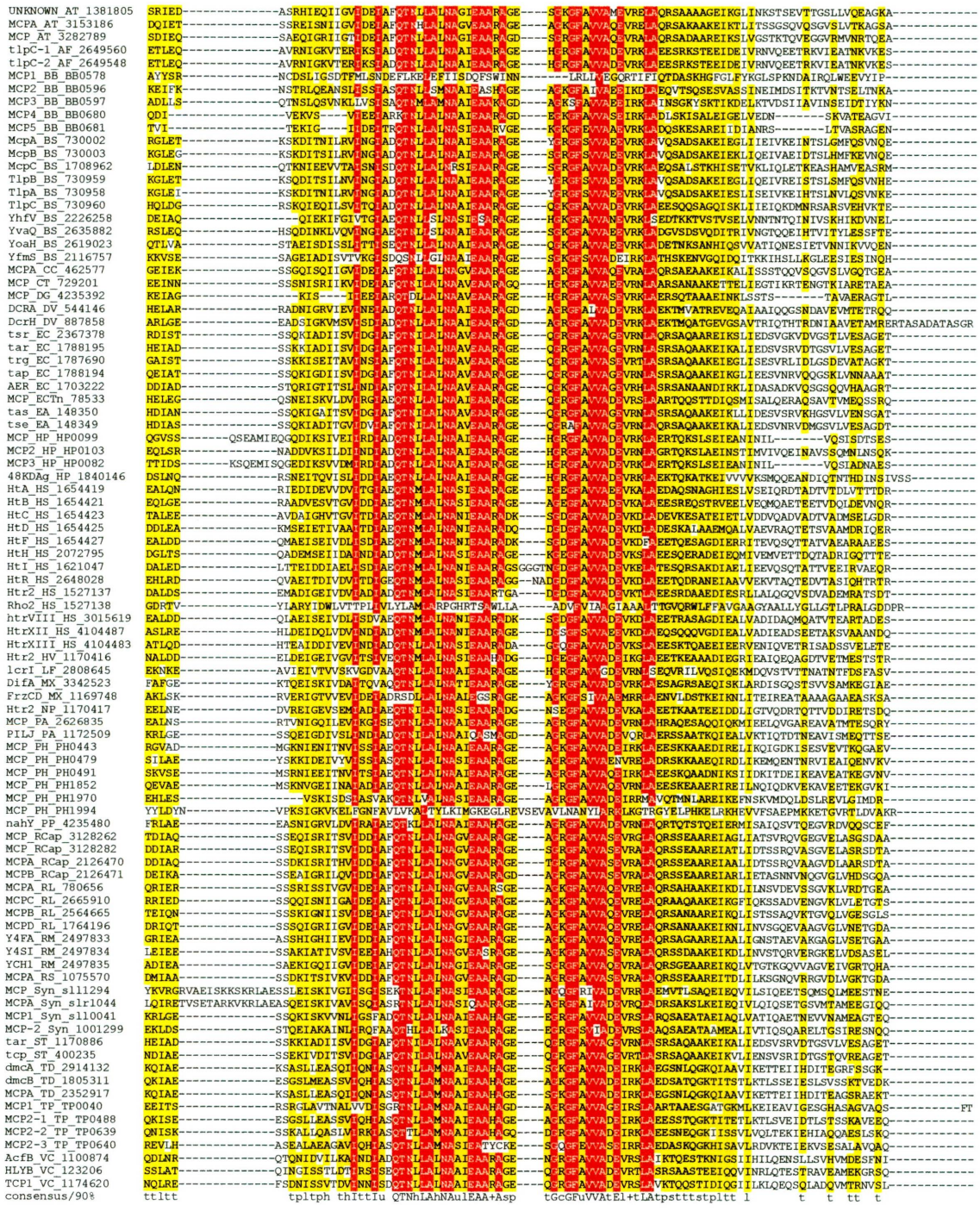
C. C-terminal signaling domain analysis

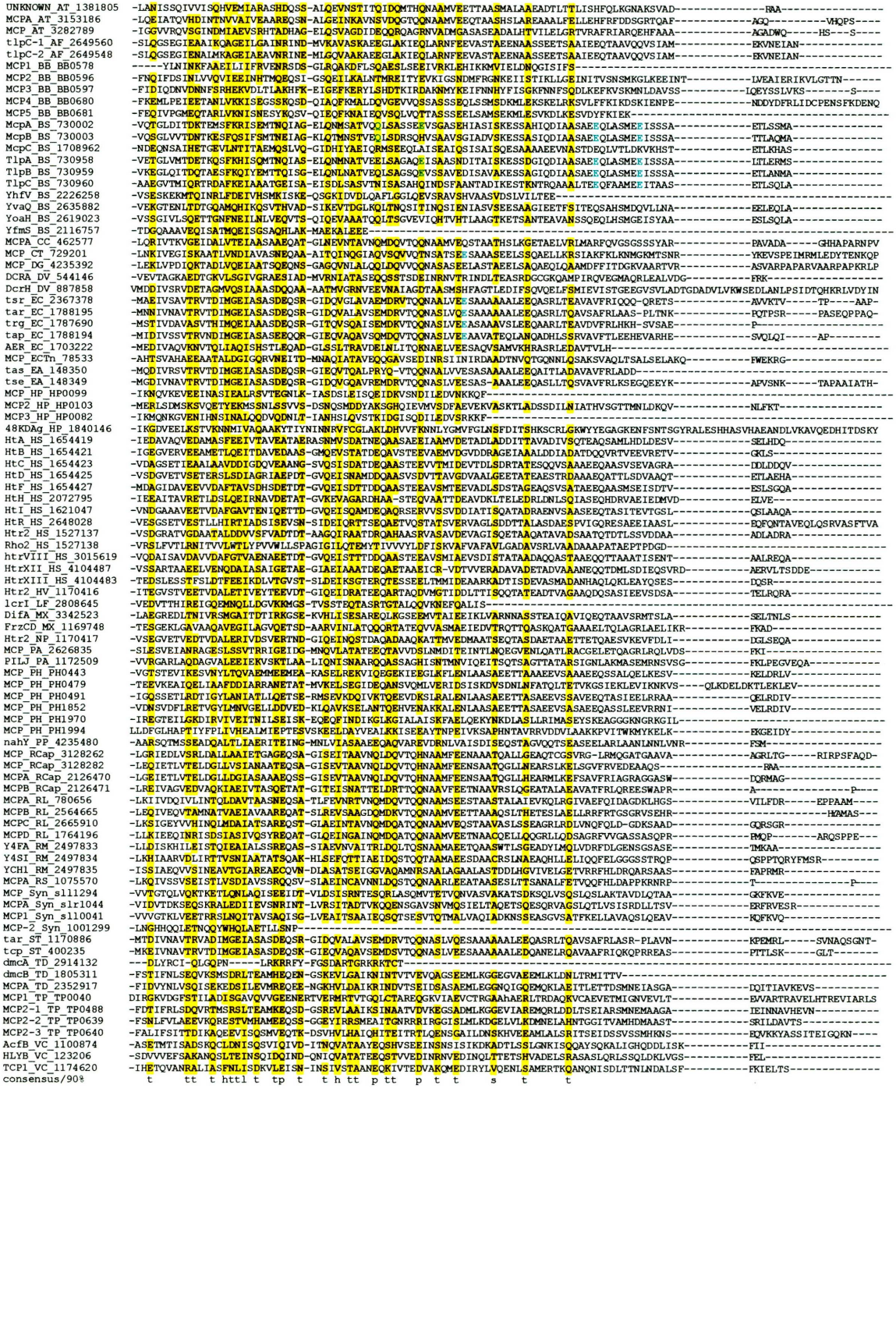
While there is great variation in the topology of MCPs, they all share the highly conserved signaling domain located in the C-terminus. This is the region that interacts with the excitation pathway. Overall, the generalized fold of the signaling domain appears to be well conserved. The best conserved regions are the two methylation domains (K1, residues 264-332 and R1, residues 438-514) and the highly conserved domain (HCD, residues 348-427). We were able to derive the consensus Q-T-N-h-L-A-h-N-A-u-l-E-A-A-+-A-s-p-t-G-c-G-F-u-V-V-A-t-E-l-+-t-L-A for the HCD (where h = hydrophobic residues, l = aliphatic residues, p = polar residues, s = small residues, t = turnlike residues, u = tiny residues and + = positively charged residues) starting at Q366 in the reference sequence Tar from *E. coli* (Figure 12).

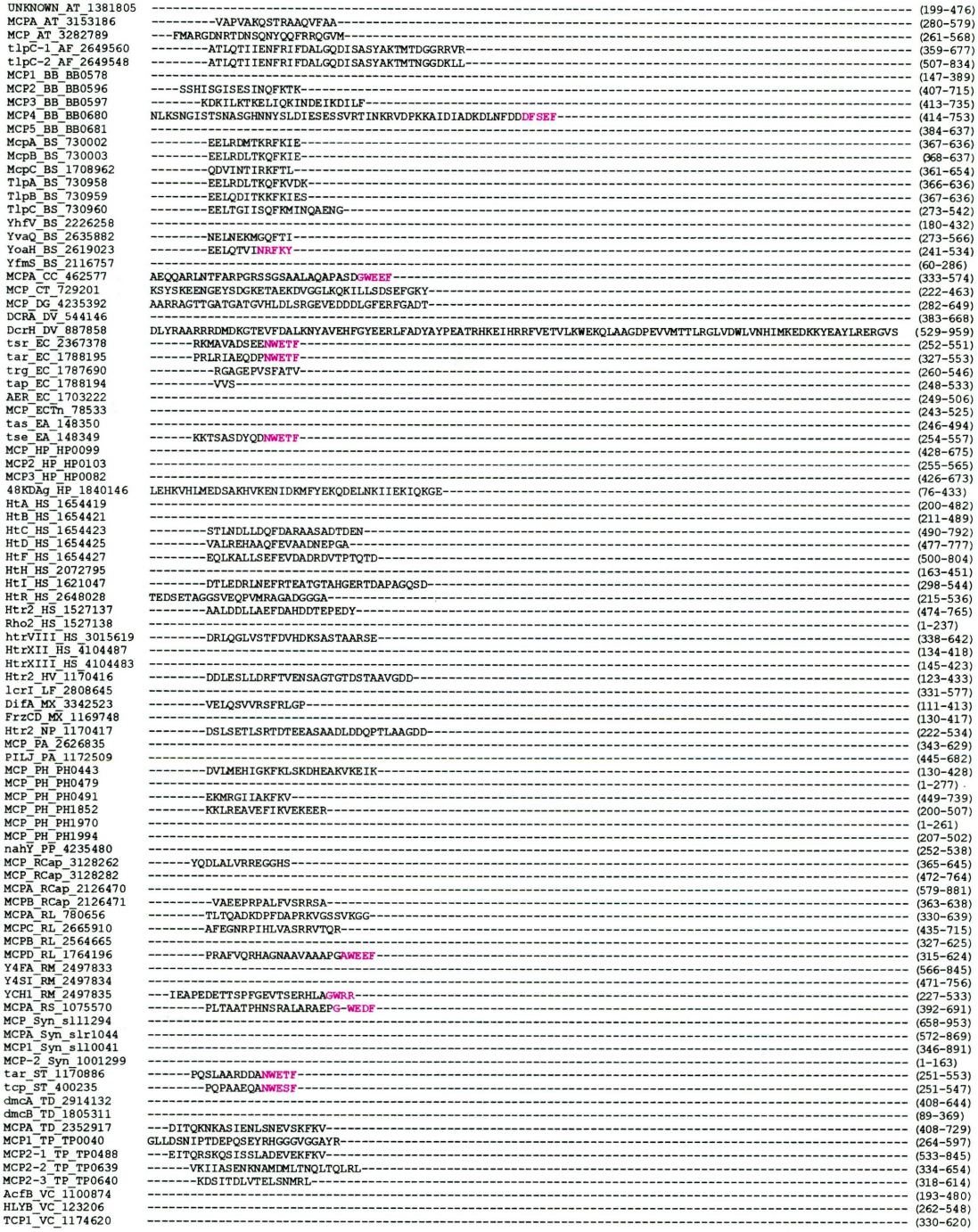
Interestingly, there are three sequences in the alignment that lack the high degree of conservation in the HCD region when compared with the other receptors. Two of them, MCP1 from *Borrelia burgdorferi* (accession number BB0578) and MCP from *Pyrococcus horikoshii* (accession number PH1994) were annotated through completely sequenced genome projects. Our alignment shows a lack of conservation, suggesting that these proteins are either non-functional or represent a deviation from the currently accepted consensus. Similarly, the sensory rhodopsin II (Rho2) from *H. salinarum* (accession number 1527138) also lacks a high degree of similarity in the HCD region. Rho2 has been purposed to be a part of the most ancient receptor-transducer complex

Figure 12. Multiple alignment of methyl-accepting chemotaxis proteins. Sequences were retrieved through BLAST (Altschul et al., 1997) searches of non-redundant databases using various MCP sequences as bait. Sequences were aligned using Clustal X (Jeanmougin et al., 1998; Thompson et al., 1997) and manual alignment. Similar amino acid residues are highlighted in yellow; identical residues are highlighted in red. Residues were colored according to a 90% consensus (generated at http://www.bork.embl-heidelberg.de/Alignment/consensus.html [N. Brown and J. Lai, unpublished]); +, positively charged residues (H, K and R); h, hydrophobic residues (A, C, F, I, L, M, V, W and Y); l, aliphatic residues (I, L and V); p, polar residues (C, D, E, H, K, N, Q, R, S and T); s, small residues (A, C, D, G, N, P, S, T and V); t, turnlike residues (A, C, D, E, G, H, K, N, Q, R, S and T; u, tiny residues (A, G and S). Residues in light-blue represent identified methylation residues; residues in pink represent known and putative CheR binding sites. The left column indicates protein/gene name, species abbreviation and accession numbers separated by underscores. All accession numbers are GenBank designations except those that start with BB, HP, PH and TP (TIGR website - http://www.tigr.org/tdb) or slr and sll (CyanoBase website - http://www.kazusa.or.jp/cyano/cyano.html). See Table 4 for definition of species abbreviations.


UNKNOWN_AT_1381805
MCPA_AT_3153186
MCP_AT_3282789
tlpC-1_AF_2649560
tlpC-2_AF_2649548
MCP1_BB_BB0578
MCP2_BB_BB0596
MCP3_BB_BB0597
MCP4_BB_BB0680
MCP5_BB_BB0681
McpA_BS_730002
McpB_BS_730003
McpC_BS_1708962
TlpA_BS_730958
TlpB_BS_730959
TlpC_BS_730960
YhfV_BS_2226258
YvaQ_BS_2635882
YoaH_BS_2619023
YfmS_BS_2116757
MCPA_CC_462577
MCP_CT_729201
MCP_DG_4235392
DCRA_DV_544146
DcrH_DV_887858
tsr_EC_2367378
tar_EC_1788195
trg_EC_1787690
tap_EC_1788194
AER_EC_1703222
MCP_ECTn_78533
tas_EA_148350
tse_EA_148349
MCP_HP_HP0099
MCP2_HP_HP0103
MCP3_HP_HP0082
48KDAg_HP_1840146
HtA_HS_1654419
HtB_HS_1654421
HtC_HS_1654423
HtD_HS_1654425
HtF_HS_1654427
HtH_HS_2072795
HtI_HS_1621047
HtR_HS_2648028
Htr2_HS_1527137
Rho2_HS_1527138
htrVIII_HS_3015619
HtrXII_HS_4104487
HtrXIII_HS_4104483
Htr2_HV_1170416
lcrI_LF_2808645
DifA_MX_3342523
FrzCD_MX_1169748
Htr2_NP_1170417
MCP_PA_2626835
PilJ_PA_1172509
MCP_PH_PH0443
MCP_PH_PH0479
MCP_PH_PH0491
MCP_PH_PH1852
MCP_PH_PH1970
MCP_PH_PH1994
nahY_PP_4235480
MCPA_RCap_2126470
MCPB_RCap_2126471
MCP_RCap_3128262
MCP_RCap_3128282
MCPA_RL_780656
MCPB_RL_2564665
MCPC_RL_2665910
MCPD_RL_1764196
Y4FA_RM_2497833
Y4SI_RM_2497834
YCH1_RM_2497835
MCPA_RS_1075570
MCP_Syn_s111294
MCPA_Syn_slr1044
MCP1_Syn_s110041
MCP-2_Syn_1001299
tar_ST_1170886
tcp_ST_400235
dmcA_TD_2914132
dmcB_TD_1805311
MCPA_TD_2352917
MCP1_TP_TP0040
MCP2-1_TP_TP0488
MCP2-2_TP_TP0639
MCP2-3_TP_TP0640
AcfB_VC_1100874
HLYB_VC_123206
TCP1_VC_1174620
consensus/90%


UNKNOWN_AT_1381805
MCPA_AT_3153186
MCP_AT_3282789
tlpC-1_AF_2649560
tlpC-2_AF_2649548
MCP1_BB_BB0578
MCP2_BB_BB0596
MCP3_BB_BB0597
MCP4_BB_BB0680
MCP5_BB_BB0681
McpA_BS_730002
McpB_BS_730003
McpC_BS_1708962
TlpB_BS_730959
TlpA_BS_730958
TlpC_BS_730960
YhfV_BS_2226258
YvaQ_BS_2635882
YoaH_BS_2619023
YfmS_BS_2116757
MCPA_CC_462577
MCP_CT_729201
MCP_DG_4235392
DCRA_DV_544146
DcrH_DV_887858
tsr_EC_2367378
tar_EC_1788195
trg_EC_1787690
tap_EC_1788194
AER_EC_1703222
MCP_ECTn_78533
tas_EA_148350
tse_EA_148349
MCP_HP_HP0099
MCP2_HP_HP0103
MCP3_HP_HP0082
48KDAg_HP_1840146
HtA_HS_1654419
HtB_HS_1654421
HtC_HS_1654423
HtD_HS_1654425
HtF_HS_1654427
HtH_HS_2072795
HtI_HS_1621047
HtR_HS_2648028
Htr2_HS_1527137
Rho2_HS_1527138
htrVIII_HS_3015619
HtrXII_HS_4104487
HtrXIII_HS_4104483
Htr2_HV_1170416
lcrI_LF_2808645
DifA_MX_3342523
FrzCD_MX_1169748
Htr2_NP_1170417
MCP_PA_2626835
PILJ_PA_1172509
MCP_PH_PH0443
MCP_PH_PH0479
MCP_PH_PH0491
MCP_PH_PH1852
MCP_PH_PH1970
MCP_PH_PH1994
nahY_PP_4235480
MCP_RCap_3128262
MCP_RCap_3128282
MCPA_RCap_2126470
MCPB_RCap_2126471
MCPA_RL_780656
MCPC_RL_2665910
MCPB_RL_2564665
MCPD_RL_1764196
Y4FA_RM_2497833
Y4SI_RM_2497834
YCH1_RM_2497835
MCPA_RS_1075570
MCP_Syn_s111294
MCPA_Syn_slr1044
MCP1_Syn_s110041
MCP-2_Syn_1001299
tar_ST_1170886
tcp_ST_400235
dmcA_TD_2914132
dmcB_TD_1805311
MCPA_TD_2352917
MCP1_TP_TP0040
MCP2-1_TP_TP0488
MCP2-2_TP_TP0639
MCP2-3_TP_TP0640
AcfB_VC_1100874
HLYB_VC_123206
TCP1_VC_1174620
consensus/90%


UNKNOWN_AT_1381805
MCPA_AT_3153186
MCP_AT_3282789
tlpC-1_AF_2649560
tlpC-2_AF_2649548
MCP1_BB_BB0578
MCP2_BB_BB0596
MCP3_BB_BB0597
MCP4_BB_BB0680
MCP5_BB_BB0681
McpA_BS_730002
McpB_BS_730003
McpC_BS_1708962
TlpA_BS_730958
TlpB_BS_730959
TlpC_BS_730960
YhfV_BS_2226258
YvaQ_BS_2635882
YoaH_BS_2619023
YfmS_BS_2116757
MCPA_CC_462577
MCP_CT_729201
MCP_DG_4235392
DCRA_DV_544146
DcrH_DV_887858
tsr_EC_2367378
tar_EC_1788195
trg_EC_1787690
tap_EC_1788194
AER_EC_1703222
MCP_ECTn_78533
tas_EA_148350
tse_EA_148349
MCP_HP_HP0099
MCP2_HP_HP0103
MCP3_HP_HP0082
48KDAg_HP_1840146
HtA_HS_1654419
HtB_HS_1654421
HtC_HS_1654423
HtD_HS_1654425
HtF_HS_1654427
HtH_HS_2072795
HtI_HS_1621047
HtR_HS_2648028
Htr2_HS_1527137
Rho2_HS_1527138
htrVIII_HS_3015619
HtrXII_HS_4104487
HtrXIII_HS_4104483
Htr2_HV_1170416
lcrI_LF_2808645
DifA_MX_3342523
FrzCD_MX_1169748
Htr2_NP_1170417
MCP_PA_2626835
PILJ_PA_1172509
MCP_PH_PH0443
MCP_PH_PH0479
MCP_PH_PH0491
MCP_PH_PH1852
MCP_PH_PH1970
MCP_PH_PH1994
nahY_PP_4235480
MCP_RCap_3128262
MCP_RCap_3128282
MCPA_RCap_2126470
MCPB_RCap_2126471
MCPA_RL_780656
MCPB_RL_2564665
MCPC_RL_2665910
MCPD_RL_1764196
Y4FA_RM_2497833
Y4SI_RM_2497834
YCH1_RM_2497835
MCPA_RS_1075570
MCP_Syn_sll1294
MCPA_Syn_slr1044
MCP1_Syn_sll0041
MCP-2_Syn_1001299
tar_ST_1170886
tcp_ST_400235
dmcA_TD_2914132
dmcB_TD_1805311
MCPA_TD_2352917
MCP1_TP_TP0040
MCP2-1_TP_TP0488
MCP2-2_TP_TP0639
MCP2-3_TP_TP0640
AcfB_VC_1100874
HLYB_VC_123206
TCP1_VC_1174620
consensus/90%


UNKNOWN_AT_1381805 (199-476)
MCPA_AT_3153186 (280-579)
MCP_AT_3282789 (261-568)
tlpC-1_AF_2649560 (359-677)
tlpC-2_AF_2649548 (507-834)
MCP1_BB_BB0578 (147-389)
MCP2_BB_BB0596 (407-715)
MCP3_BB_BB0597 (413-735)
MCP4_BB_BB0680 (414-753)
MCP5_BB_BB0681 (384-637)
McpA_BS_730002 (367-636)
McpB_BS_730003 (368-637)
McpC_BS_1708962 (361-654)
TlpA_BS_730958 (366-636)
TlpB_BS_730959 (367-636)
TlpC_BS_730960 (273-542)
YhfV_BS_2226258 (180-432)
YvaQ_BS_2635882 (273-566)
YoaH_BS_2619023 (241-534)
YfmS_BS_2116757 (60-286)
MCPA_CC_462577 (333-574)
MCP_CT_729201 (222-463)
MCP_DG_4235392 (282-649)
DCRA_DV_544146 (383-668)
DcrH_DV_887858 (529-959)
tsr_EC_2367378 (252-551)
tar_EC_1788195 (327-553)
trg_EC_1787690 (260-546)
tap_EC_1788194 (248-533)
AER_EC_1703222 (249-506)
MCP_ECTn_78533 (243-525)
tas_EA_148350 (246-494)
tse_EA_148349 (254-557)
MCP_HP_HP0099 (428-675)
MCP2_HP_HP0103 (255-565)
MCP3_HP_HP0082 (426-673)
48KDAg_HP_1840146 (76-433)
HtA_HS_1654419 (200-482)
HtB_HS_1654421 (211-489)
HtC_HS_1654423 (490-792)
HtD_HS_1654425 (477-777)
HtF_HS_1654427 (500-804)
HtH_HS_2072795 (163-451)
HtI_HS_1621047 (298-544)
HtR_HS_2648028 (215-536)
Htr2_HS_1527137 (474-765)
Rho2_HS_1527138 (1-237)
htrVIII_HS_3015619 (338-642)
HtrXII_HS_4104487 (134-418)
HtrXIII_HS_4104483 (145-423)
Htr2_HV_1170416 (123-433)
lcrI_LF_2808645 (331-577)
DifA_MX_3342523 (111-413)
FrzCD_MX_1169748 (130-417)
Htr2_NP_1170417 (222-534)
MCP_PA_2626835 (343-629)
PILJ_PA_1172509 (445-682)
MCP_PH_PH0443 (130-428)
MCP_PH_PH0479 (1-277)
MCP_PH_PH0491 (449-739)
MCP_PH_PH1852 (200-507)
MCP_PH_PH1970 (1-261)
MCP_PH_PH1994 (207-502)
nahY_PP_4235480 (252-538)
MCP_RCap_3128262 (365-645)
MCP_RCap_3128282 (472-764)
MCPA_RCap_2126470 (579-881)
MCPB_RCap_2126471 (363-638)
MCPA_RL_780656 (330-639)
MCPC_RL_2665910 (435-715)
MCPB_RL_2564665 (327-625)
MCPD_RL_1764196 (315-624)
Y4FA_RM_2497833 (566-845)
Y4SI_RM_2497834 (471-756)
YCH1_RM_2497835 (227-533)
MCPA_RS_1075570 (392-691)
MCP_Syn_s111294 (658-953)
MCPA_Syn_slr1044 (572-869)
MCP1_Syn_s110041 (346-891)
MCP-2_Syn_1001299 (1-163)
tar_ST_1170886 (251-553)
tcp_ST_400235 (251-547)
dmcA_TD_2914132 (408-644)
dmcB_TD_1805311 (89-369)
MCPA_TD_2352917 (408-729)
MCP1_TP_TP0040 (264-597)
MCP2-1_TP_TP0488 (533-845)
MCP2-2_TP_TP0639 (334-654)
MCP2-3_TP_TP0640 (318-614)
AcfB_VC_1100874 (193-480)
HLYB_VC_123206 (262-548)
TCP1_VC_1174620 (330-620)

that has a dual function: the capability to transmit both light and chemical signals (Zhang et al., 1996). Thus, Rho2, BB0578 and PH1994 could represent ancient precursors to the modern HCD.

Insertion/deletion elements (indels) similar to those previously identified (Le Moual and Koshland, 1996) were observed in the alignment, with the main differences being the location of the indels and the introduction of two new cases: one near the very beginning of the K1 region and another in the middle of the HCD. Indels are annotated "250.14+" according to LeMoual and Koshland (1996) in which 250 represents the residue number before the indel, 14 represents the length of the indel and + or - represents an insertion or deletion, respectively. The four main indels were found at the beginning of the K1 methylation region (295.13+), the end of the R1 methylation region (523.12+) and before and after the HCD (354.14+ and 439.13+). A new six-residue indel was discovered at the very beginning of the K1 region (277.6+). Interestingly a four-residue indel (379.4+) was identified right in the middle of the HCD region. However, this indel appears in only three sequences and in one case involves the insertion of only two amino acids. It is likely that these short insertions cause only minimal disruption of the predicted α-helical secondary structure. Supporting this view is the observation that one MCP with an insertion at this location (HtI from *Halobacterium salinarium*) is functional experimentally (Zhang et al., 1996).

Whereas most of the C-terminus is extremely conserved, there is one area that has a high degree of variability. This area lies in the tail end of the MCPs after the 523.12+ indel. In addition to poor sequence conservation (Figure 12) this region varies

greatly in length. For example, MCPA from *Rhodobacter capsulatus* has a very short tail (~5 amino acids), while others, such as DcrH from *Desulfovibrio vulgaris*, are much longer (~150 amino acids). Most likely the presence of these long tails on certain receptors reflects an additional functional site. One known functional site is the binding motif for the methyltransferase CheR. Located at the extreme end of some receptors, such as Tsr and Tar from *E. coli*, the motif is comprised of a conserved pentapeptide sequence N-W-E-T/S-F. Not all receptors contain this motif and the current hypothesis states that those receptors that do contain this motif and bind CheR hold it in close proximity to those receptors that lack the motif (Wu et al., 1996; Okumura et al., 1998). Analysis of the alignment reveals the presence of the same or similar motifs located at the end of a few other receptors (Figure 12). It is not unexpected to find only a few receptors with this motif. Some genomes do not contain a CheR protein (Zhulin, 1999) and therefore receptors from these genomes lack the CheR binding motif. Others may be modified in a different manner (i.e. covalent modification) or not methylated at all. This doesn't preclude the possibility of a functionally similar motif with a different peptide sequence located on the other receptors.

Those receptors that are methylated contain the consensus E/Q-E/Q-X-X-A-S/T where X represents any amino acid and the underlined E/Q represents the site of modification (Terwilliger et al., 1986). Methylation sites are identified on the alignment (Figure 12) and a few E/Q residues correspond to methylation sites identified experimentally (Le Moual and Koshland, 1996). However it is complicated to determine actual methylation sites based on alignments alone since many receptors may be

methylated in different regions, have a different methylation consensus or not be methylated. Thus experiments should be performed to determine the actual site of methylation using the alignment as a guide.

D. N-terminal sensing domain analysis

Since the N-terminal portion of receptors can be highly variable, we also analyzed this domain of the identified MCPs. A PSI-BLAST (Altschul et al., 1997) search was performed using the N-terminal from each MCP as a query. All MCPs demonstrated a statistically significant relationship to other MCPs ($E < .001$), and a majority of them revealed a significant relationship to sensor kinases ($E < .001$), specifically histidine and a few serine/threonine kinases (Table 6). Simulations suggest that nonhomologous swapping of DNA segments that encode individual protein structural elements may be a means for creating new proteins (Bogarad and Deem, 1999; Henikoff et al., 1997). Based on this, it appears that in the past a kinase-sensing domain "fused" with a conserved signaling domain to form a new receptor. Therefore we hypothesize that the N-terminal sensing domain and the conserved C-terminal signaling domain of bacterial chemotaxis receptors were at one time separate elements that eventually "fused" to create a functional receptor. Analysis of chemotaxis systems in completely sequenced genomes helps to illustrate and support our hypothesis. In one of the *Synechocystis* operons the conserved C-terminal signaling domain is duplicated, while the N-terminal sensing domain appears by itself in one of the chemotaxis operons of *Borrelia burgdorferi* (Zhulin, 1999).

Table 6. Examples of kinases retrieved from N-terminal PSI-BLAST searches.[a]

Class	MCP N-terminus query	Kinases retrieved and E-values
Large periplasmic region	YvaQ (2635882) from *Bacillus subtilis*	histdine kinase (4104609) from *Lactobacillus sakei* ($E=2x10^{-22}$)
		histidine kinase (2352098) from *Pseudomonas aeruginosa* ($E=9x10^{-21}$)
		histidine kinase (4336932) from *Nostoc punctiforme* ($E=2x10^{-20}$)
	MCP2 (BB0596) from *Borrelia burgdorferi*	Serine/threonine kinase (3845109) from *Plasmodium falciparum* ($E=2x10^{-30}$)
No periplasmic region	YoaH (2619023) from *Bacillus subtilis*	histidine kinase (4104603) from *Lactobacillus sakei* ($E=5x10^{-33}$)
		histidine kinase (4336932) from *Nostoc punctiforme* ($E=2x10^{-28}$)
		histidine kinase (2352098) from *Pseudomonas aeruginosa* ($E=2x10^{-22}$)
	Unknown (1381805) from *Agrobacterium tumefaciens*	histidine kinase (1653308) from *Synechocystis* ($E=5x10^{-16}$)
		histidine kinase (2739133) from *Myxococcus xanthus* ($E=5x10^{-16}$)
Small periplasmic region	Tn1721(78533) from *Escherichia coli*	histidine kinase (2650219) from *Archaeoglobus fulgidus* ($E = 2x10^{-30}$)
		histidine kinase (2621877) from *Methanobacterium thermoautotrophicum* ($E=3x10^{-30}$)
Cytoplasmic sensor	MCPA (1075570) from *Rhodobacter sphaeroides*	sensor kinase (777753) from *Vibrio cholerae* ($E=4x10^{-29}$)
		histidine kinase (2338728) from *Calothrix viguieri* ($E=2x10^{-20}$)

[a]Numbers in parentheses represent GenBank protein identification numbers. The number starting with BB is accessible via the TIGR website (http://www.tigr.org/tdb).

In summary, we have identified over 90 known and putative bacterial chemoreceptors and classified them into six distinct classes based on topology studies. The C-terminal signaling domain is highly conserved, both structurally (generalized fold) and functionally (HCD interaction with the chemotaxis pathway). Finally, we have proposed that the sensing/signal domains were once separate entities that joined together during the course of molecular evolution to form the bacterial chemotaxis receptors.

APPENDIX A

A sequence in FASTA format begins with a single-line description followed by lines of sequence data (DNA or amino acids). The description line is distinguished from the sequence data by having a greater-than ('>') placed at the beginning of the description line. It is recommended that all lines of text be shorter than 80 characters in length, however one need not always follow this guideline. Some programs will allow more characters per line, while others allow less. Below the amino acid sequence of the chemotaxis protein CheY from *E. coli* is used as an example of a sequence in FASTA format:

```
>gi|116291|sp|P06143|CHEY_ECOLI CHEMOTAXIS PROTEIN CHEY
MADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMPNMDGLELLK
TIRADGAMSALPVLMVTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLGM
```

Sequences are should be represented in the standard International Union of Pure and Applied Chemistry (IUPAC) amino and nucleic acid codes, with some exceptions. Lower-case letters are usually accepted and are mapped into upper-case; a single hyphen or dash can be used to represent gaps of indeterminate length; and in amino acid sequences, sometimes U and * are acceptable letters. Before inputting you data, any numerical digits in the query sequence should either be removed or replaced by appropriate letter codes.

The nucleic acid codes that can be supported by different programs are:

```
A --> adenosine          M --> A C (amino)
C --> cytidine           S --> G C (strong)
G --> guanine            W --> A T (weak)
T --> thymidine          B --> G T C
```

```
U --> uridine              D --> G A T
R --> G A (purine)         H --> A C T
Y --> T C (pyrimidine)     V --> G C A
K --> G T (keto)           N --> A G C T (any)
                           -  gap of indeterminate length
```

The one letter amino acid codes that are most commonly accepted are:

```
A  alanine                          P  proline
B  aspartate or asparagine          Q  glutamine
C  cysteine                         R  arginine
D  aspartate                        S  serine
E  glutamate                        T  threonine
F  phenylalanine                    U  selenocysteine
G  glycine                          V  valine
H  histidine                        W  tryptophan
I  isoleucine                       Y  tyrosine
K  lysine                           Z  glutamate or glutamine
L  leucine                          X  any
M  methionine                       *  translation stop
N  asparagine                       -  gap of indeterminate length
```

It should also be noted that when aligning multiple sequences, the font "courier new" is used. This font is used so that the width of each individual character typed is the same. This is helpful for lining up residues in multiple alignments, but it is not necessary to have one's input in this font.

REFERENCES

Alex, L.A., Borkovich, K.A., and Simon, M.I. (1996). Hyphal development in *Neurospora crassa*: involvement of a two-component histidine kinase. Proc.Natl.Acad.Sci.U.S.A. *93*, 3416-3421.

Altschul, S.F. (1998). Fundamentals of database searching. In Trends Guide to Bioinformatics. S. Brenner and F. Lewitter, eds. (Oxford: Elsevier Science Ltd.), pp. 7-9.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J.Mol.Biol. *215* , 403-410.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic.Acids.Res. *25*, 3389-3402.

Ames, P., Chen, J., Wolff, C., and Parkinson, J.S. (1988). Structure-function studies of bacterial chemosensors. Cold Spring Harb.Symp.Quant.Biol. *53 Pt 1:59-65*, 59-65.

Appleby, J.L. and Bourret, R.B. (1999). Activation of CheY mutant D57N by phosphorylation at an alternate site, Ser56. Submitted

Appleby, J.L., Parkinson, J.S., and Bourret, R.B. (1996). Signal transduction via the multi-step phosphorelay: not necessarily a road less traveled. Cell *86*, 845-848.

Barak, R. and Eisenbach, M. (1992). Correlation between phosphorylation of the chemotaxis protein CheY and its activity at the flagellar motor [published erratum appears in Biochemistry 1992 May 19;31(19):4736]. Biochemistry *31*, 1821-1826.

Baralle, F.E. (1977). Complete nucleotide sequence of the 5' noncoding region of rabbit beta- globin mRNA. Cell *10*, 549-558.

Barrell, B.G. and Clarck, B.F.C. (1974). Handbook of Nucleic Acid Sequences (Oxford: Joynson-Bruvvers Ltd.).

Bellsolell, L., Prieto, J., Serrano, L., and Coll, M. (1994). Magnesium binding to the bacterial chemotaxis protein CheY results in large conformational changes involving its functional surface [published erratum appears in J Mol Biol 1994 Sep 9;242(1):103]. J.Mol.Biol *238*, 489-495.

Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F., Rapp, B.A., and Wheeler, D.L. (1999). GenBank. Nucleic.Acids.Res. *27*, 12-17.

Berg, H.C. (1993). Random Walks in Biology (Princeton: Princeton University Press).

Berg, H.C. and Anderson, R.A. (1973). Bacteria swim by rotating their flagellar filaments. Nature *245*, 380-382.

Berg, H.C. and Brown, D.A. (1972). Chemotaxis in *Escherichia coli* analysed by three-dimensional tracking. Nature *239*, 500-504.

Bibikov, S.I., Biran, R., Rudd, K.E., and Parkinson, J.S. (1997). A signal transducer for aerotaxis in *Escherichia coli*. J.Bacteriol. *179*, 4075-4079.

Bilwes, A.M., Alex, L.A., Crane, B.R., and Simon, M.I. (1999). Structure of CheA, a signal-transducing histidine kinase. Cell *96*, 131-141.

Blair, D.F. (1995). How bacteria sense and swim. Annu.Rev.Microbiol. *49:489-522*, 489-522.

Blat, Y. and Eisenbach, M. (1996). Conserved C-terminus of the phosphatase CheZ is a binding domain for the chemotactic response regulator CheY. Biochemistry *35*, 5679-5683.

Bogarad, L.D. and Deem, M.W. (1999). A hierarchical approach to protein molecular evolution. Proc.Natl.Acad.Sci.U.S.A. *96*, 2591-2595.

Boguski, M.S. (1998). Bioinformatics - a new era. In Trends Guide to Bioinformatics. S. Brenner and F. Lewitter, eds. (Oxford: Elsevier Science Ltd.), pp. 1-3.

Borkovich, K.A., Kaplan, N., Hess, J.F., and Simon, M.I. (1989). Transmembrane signal transduction in bacterial chemotaxis involves ligand-dependent activation of phosphate group transfer. Proc.Natl.Acad.Sci.U.S.A. *86*, 1208-1212.

Bourret, R.B., Davagnino, J., and Simon, M.I. (1993). The carboxy-terminal portion of the CheA kinase mediates regulation of autophosphorylation by transducer and CheW. J.Bacteriol. *175*, 2097-2101.

Boyd, A., Krikos, A., and Simon, M. (1981). Sensory transducers of *E. coli* are encoded by homologous genes. Cell *26*, 333-343.

Bren, A. and Eisenbach, M. (1998). The N terminus of the flagellar switch protein, FliM, is the binding domain for the chemotactic response regulator, CheY. J.Mol.Biol. *278*, 507-514.

Brenner, S.E. (1999). Errors in genome annotation. Trends Genet. *15*, 132-133.

Chervitz, S.A. and Falke, J.J. (1996). Molecular mechanism of transmembrane signaling by the aspartate receptor: a model. Proc.Natl.Acad.Sci.U.S.A. *93*, 2545-2550.

Chervitz, S.A., Lin, C.M., and Falke, J.J. (1995). Transmembrane signaling by the aspartate receptor: engineered disulfides reveal static regions of the subunit interface. Biochemistry *34*, 9722-9733.

Danielson, M.A., Bass, R.B., and Falke, J.J. (1997). Cysteine and disulfide scanning reveals a regulatory alpha-helix in the cytoplasmic domain of the aspartate receptor. J.Biol.Chem. *272*, 32878-32888.

Dayhoff, M.O. (1969). Computer analysis of protein evolution. Sci.Am. *221*, 86-95.

Deakin, W.J., Sanderson, J.L., Goswami, T., and Shaw, C.H. (1997). The *Agrobacterium tumefaciens* motor gene, motA, is in a linked cluster with the flagellar switch protein genes, fliG, fliM and fliN. Gene *189*, 139-141.

DeRosier, D.J. (1998). The turn of the screw: the bacterial flagellar motor. Cell *93*, 17-20.

Ditty, J.L., Grimm, A.C., and Harwood, C.S. (1998). Identification of a chemotaxis gene region from *Pseudomonas putida*. FEMS Microbiol.Lett. *159*, 267-273.

Djordjevic, S., Goudreau, P.N., Xu, Q., Stock, A.M., and West, A.H. (1998). Structural basis for methylesterase CheB regulation by a phosphorylation-activated domain. Proc.Natl.Acad.Sci.U.S.A. *95*, 1381-1386.

Doolittle, R.F. (1997). Some reflections on the early days of sequence searching. J.Mol.Med. *75*, 239-241.

Efstratiadis, A., Kafatos, F.C., and Maniatis, T. (1977). The primary structure of rabbit beta-globin mRNA as determined from cloned DNA. Cell *10*, 571-585.

Falke, J.J., Bass, R.B., Butler, S.L., Chervitz, S.A., and Danielson, M.A. (1997). The two-component signaling pathway of bacterial chemotaxis: a molecular view of signal transduction by receptors, kinases, and adaptation enzymes. Annu.Rev.Cell Dev.Biol *13:457-512*, 457-512.

Felsenstein, J. (1989). *PHYLIP -- Phylogeny Inference Package (Version 3.2)*. Cladistics 5: 164-166.

Fitch, W.M. (1966). Evidence suggesting a partial, internal duplication in the ancestral gene for heme-containing globins. J.Mol.Biol. *16*, 17-27.

Fitch, W.M. and Margoliash, E. (1967). Construction of phylogenetic trees. Science *155*, 279-284.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., and Merrick, J.M. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science *269*, 496-512.

Franklin, J. (1991). The Future of the Medical Journal. S.P. Lock, ed. (London: BMJ Press), pp. 400-403.

Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., Gwinn, M., Dougherty, B., Tomb, J.F., Fleischmann, R.D., Richardson, D., Peterson, J., Kerlavage, A.R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M.D., Gocayne, J., and Venter, J.C. (1997). Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature *390*, 580-586.

Frishman, D. and Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. Protein Eng. *9*, 133-142.

Frishman, D. and Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. Proteins *27*, 329-335.

Gardina, P.J. and Manson, M.D. (1996). Attractant signaling by an aspartate chemoreceptor dimer with a single cytoplasmic domain. Science *274*, 425-426.

Grebe, T.W. and Stock, J. (1998). Bacterial chemotaxis: the five sensors of a bacterium. Curr.Biol. *8*, R154-R157

Hazelbauer, G.L. and Adler, J. (1971). Role of the galactose binding protein in chemotaxis of *Escherichia coli* toward galactose. Nat.New Biol. *230*, 101-104.

Henikoff, S., Greene, E.A., Pietrokovski, S., Bork, P., Attwood, T.K., and Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. Science *278*, 609-614.

Hess, J.F., Bourret, R.B., and Simon, M.I. (1988). Histidine phosphorylation and phosphoryl group transfer in bacterial chemotaxis. Nature *336*, 139-143.

Hess, J.F., Oosawa, K., Kaplan, N., and Simon, M.I. (1988). Phosphorylation of three proteins in the signaling pathway of bacterial chemotaxis. Cell *53*, 79-87.

Ishikawa, J. and Hotta, K. (1999). FramePlot: a new implementation of the Frame analysis for predicting protein-coding regions in bacterial DNA with a high G+C content. FEMS Microbiol.Lett. 251-253.

Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., and Gibson, T.J. (1998). Multiple sequence alignment with Clustal X. Trends.Biochem.Sci. *23*, 403-405.

Kehry, M.R. and Dahlquist, F.W. (1982). The methyl-accepting chemotaxis proteins of *Escherichia coli*. Identification of the multiple methylation sites on methyl-accepting chemotaxis protein I. J.Biol.Chem. *257*, 10378-10386.

Kidwell, P.A. and Ceruzzi, P.E. (1999). Landmarks in Digital Computing (Washington, D.C.: Smithsonian Institution Press).

Koonin, E. (1999). Why genome analysis? Trends Genet. *15*, 131

Kort, E.N., Goy, M.F., Larsen, S.H., and Adler, J. (1975). Methylation of a membrane protein involved in bacterial chemotaxis. Proc.Natl.Acad.Sci.U.S.A. *72*, 3939-3943.

Kuo, S.C. and Koshland, D.E.J. (1987). Roles of cheY and cheZ gene products in controlling flagellar rotation in bacterial chemotaxis of *Escherichia coli*. J.Bacteriol. *169*, 1307-1314.

Le Moual, H. and Koshland, D.E.J. (1996). Molecular evolution of the C-terminal cytoplasmic domain of a superfamily of bacterial receptors involved in taxis. J.Mol.Biol. *261*, 568-585.

Li, J., Swanson, R.V., Simon, M.I., and Weis, R.M. (1995). The response regulators CheB and CheY exhibit competitive binding to the kinase CheA. Biochemistry *34*, 14626-14636.

Lukashin, A.V. and Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic.Acids.Res. *26*, 1107-1115.

Lupas, A. and Stock, J. (1989). Phosphorylation of an N-terminal regulatory domain activates the CheB methylesterase in bacterial chemotaxis. J.Biol Chem. *264*, 17337-17342.

Maeda, T., Wurgler-Murphy, S.M., and Saito, H. (1994). A two-component system that regulates an osmosensing MAP kinase cascade in yeast. Nature *369*, 242-245.

Masduki, A., Nakamura, J., Ohga, T., Umezaki, R., Kato, J., and Ohtake, H. (1995). Isolation and characterization of chemotaxis mutants and genes of *Pseudomonas aeruginosa*. J.Bacteriol. *177*, 948-952.

Mathews, M.A., Tang, H.L., and Blair, D.F. (1998). Domain analysis of the FliM protein of *Escherichia coli*. J.Bacteriol. *180*, 5580-5590.

McEvoy, M.M., Bren, A., Eisenbach, M., and Dahlquist, F.W. (1999). Identification of the binding interfaces on CheY for two of its targets, the phosphatase CheZ and the flagellar switch protien FliM. J.Mol.Biol. (in press)

McEvoy, M.M., Hausrath, A.C., Randolph, G.B., Remington, S.J., and Dahlquist, F.W. (1998). Two binding modes reveal flexibility in kinase/response regulator interactions in the bacterial chemotaxis pathway. Proc.Natl.Acad.Sci.U.S.A. *95*, 7333-7338.

Milburn, M.V., Prive, G.G., Milligan, D.L., Scott, W.G., Yeh, J., Jancarik, J., Koshland, D.E.J., and Kim, S.H. (1991). Three-dimensional structures of the ligand-binding domain of the bacterial aspartate receptor with and without a ligand. Science *254*, 1342-1347.

Milligan, D.L. and Koshland, D.E.J. (1988). Site-directed cross-linking. Establishing the dimeric structure of the aspartate receptor of bacterial chemotaxis. J.Biol.Chem. *263*, 6268-6275.

Morel-Deville, F., Fauvel, F., and Morel, P. (1998). Two-component signal-transducing systems involved in stress responses and vancomycin susceptibility in *Lactobacillus sakei*. Microbiology. *144*, 2873-2883.

Moy, F.J., Lowry, D.F., Matsumura, P., Dahlquist, F.W., Krywko, J.E., and Domaille, P.J. (1994). Assignments, secondary structure, global fold, and dynamics of chemotaxis Y protein using three- and four-dimensional heteronuclear (13C,15N) NMR spectroscopy. Biochemistry *33*, 10731-10742.

Musacchio, A., Wilmanns, M., and Saraste, M. (1994). Structure and function of the SH3 domain. Prog.Biophys.Mol.Biol. *61*, 283-297.

Mutoh, N. and Simon, M.I. (1986). Nucleotide sequence corresponding to five chemotaxis genes in *Escherichia coli*. J.Bacteriol. *165*, 161-166.

Ninfa, E.G., Stock, A., Mowbray, S., and Stock, J. (1991). Reconstitution of the bacterial chemotaxis signal transduction system from purified components. J.Biol.Chem. *266*, 9764-9770.

Nowlin, D.M., Bollinger, J., and Hazelbauer, G.L. (1987). Sites of covalent modification in Trg, a sensory transducer of *Escherichia coli*. J.Biol.Chem. *262*, 6039-6045.

Okumura, H., Nishiyama, S., Sasaki, A., Homma, M., and Kawagishi, I. (1998). Chemotactic adaptation is altered by changes in the carboxy-terminal sequence conserved among the major methyl-accepting chemoreceptors. J.Bacteriol. *180*, 1862-1868.

Pfeffer, W. (1883). Locomotorische richtungsbewegungen durch chemische rieze. Berichte der Deutschen Botanischen Gesellschaft *1*, 524-533.

Proudfoot, N.J. (1977). Complete 3' noncoding region sequences of rabbit and human beta-globin messenger RNAs. Cell *10*, 559-570.

Ramakrishnan, R., Schuster, M., and Bourret, R.B. (1998). Acetylation at Lys-92 enhances signaling by the chemotaxis response regulator protein CheY. Proc.Natl.Acad.Sci.U.S.A. *95*, 4918-4923.

Rebbapragada, A., Johnson, M.S., Harding, G.P., Zuccarelli, A.J., Fletcher, H.M., Zhulin, I.B., and Taylor, B.L. (1997). The Aer protein and the serine chemoreceptor Tsr independently sense intracellular energy levels and transduce oxygen, redox, and energy signals for *Escherichia coli* behavior. Proc.Natl.Acad.Sci.U.S.A. *94*, 10541-10546.

Rost, B. and Sander, C. (1993a). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc.Natl.Acad.Sci.U.S.A. *90*, 7558-7562.

Rost, B. and Sander, C. (1993b). Prediction of protein secondary structure at better than 70% accuracy. J.Mol.Biol. *232*, 584-599.

Rost, B. and Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. Proteins *19*, 55-72.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol.Biol.Evol. *4*, 406-425.

Sanders, D.A., Gillece-Castro, B.L., Stock, A.M., Burlingame, A.L., and Koshland, D.E.J. (1989). Identification of the site of phosphorylation of the chemotaxis response regulator protein, CheY. J.Biol.Chem. *264*, 21770-21778.

Silversmith, R.E., Appleby, J.L., and Bourret, R.B. (1997). Catalytic mechanism of phosphorylation and dephosphorylation of CheY: kinetic characterization of imidazole phosphates as phosphodonors and the role of acid catalysis. Biochemistry *36*, 14965-14974.

Silversmith, R.E. and Bourret, R.B. (1999). Throwing the switch in bacterial chemotaxis. Trends.Microbiol. *7*, 16-22.

Simms, S.A., Stock, A.M., and Stock, J.B. (1987). Purification and characterization of the S-adenosylmethionine:glutamyl methyltransferase that modifies membrane chemoreceptor proteins in bacteria. J.Biol.Chem. *262*, 8537-8543.

Singh, M., Berger, B., Kim, P.S., Berger, J.M., and Cochran, A.G. (1998). Computational learning reveals coiled coil-like motifs in histidine kinase linker domains. Proc.Natl.Acad.Sci.U.S.A. *95*, 2738-2743.

Smith, T.F. (1990). The history of the genetic sequence databases. Genomics *6*, 701-707.

Sourjik, V. and Schmitt, R. (1998). Phosphotransfer between CheA, CheY1, and CheY2 in the chemotaxis signal transduction chain of *Rhizobium meliloti*. Biochemistry *37*, 2327-2335.

Springer, M.S., Goy, M.F., and Adler, J. (1979). Protein methylation in behavioural control mechanisms and in signal transduction. Nature *280*, 279-284.

Springer, W.R. and Koshland, D.E.J. (1977). Identification of a protein methyltransferase as the cheR gene product in the bacterial sensing system. Proc.Natl.Acad.Sci.U.S.A. *74*, 533-537.

Stewart, R.C. (1993). Activating and inhibitory mutations in the regulatory domain of CheB, the methylesterase in bacterial chemotaxis. J.Biol.Chem. *268*, 1921-1930.

Stock, A., Koshland, D.E.J., and Stock, J. (1985). Homologies between the *Salmonella typhimurium* CheY protein and proteins involved in the regulation of chemotaxis, membrane protein synthesis, and sporulation. Proc.Natl.Acad.Sci.U.S.A. *82*, 7989-7993.

Stock, A.M., Mottonen, J.M., Stock, J.B., and Schutt, C.E. (1989). Three-dimensional structure of CheY, the response regulator of bacterial chemotaxis. Nature *337*, 745-749.

Stock, A.M. and Stock, J.B. (1987). Purification and characterization of the CheZ protein of bacterial chemotaxis. J.Bacteriol. *169*, 3301-3311.

Stock, J.B. and Koshland, D.E.J. (1978). A protein methylesterase involved in bacterial sensing. Proc.Natl.Acad.Sci.U.S.A. *75*, 3659-3663.

Stock, J.B., Lukat, G.S., and Stock, A.M. (1991). Bacterial chemotaxis and the molecular logic of intracellular signal transduction networks. Annu.Rev.Biophys.Biophys.Chem. *20:109-36*, 109-136.

Stock, J.B. and Surette, M.G. (1996). Chemotaxis. In *Escherichia coli* and *Salmonella*: cellular and molecular biology. F.C. Neidhardt, R. Curtiss III, E.C.C. Lin, K.B. Low, B. Magasanik, and et al, eds. (Washington, DC: ASM), pp. 1103-1129.

Stoddard, B.L. and Koshland, D.E.J. (1992). Prediction of the structure of a receptor-protein complex using a binary docking method. Nature *358*, 774-776.

Storch, K.F., Rudolph, J., and Oesterhelt, D. (1999). Car: a cytoplasmic sensor responsible for arginine chemotaxis in the archaeon *Halobacterium salinarum*. EMBO J. *18*, 1146-1158.

Swanson, R.V., Lowry, D.F., Matsumura, P., McEvoy, M.M., Simon, M.I., and Dahlquist, F.W. (1995). Localized perturbations in CheY structure monitored by NMR identify a CheA binding interface. Nat.Struct.Biol. *2*, 906-910.

Swanson, R.V., Schuster, S.C., and Simon, M.I. (1993). Expression of CheA fragments which define domains encoding kinase, phosphotransfer, and CheY binding activities. Biochemistry *32*, 7623-7629.

Swofford, D. (1991). *Phylogenetic Analysis Using Parsimony (PAUP) Version 3.0d.* Illinois Natural History Survey, Champaign, IL.

Tanaka, T., Saha, S.K., Tomomori, C., Ishima, R., Liu, D., Tong, K.I., Park, H., Dutta, R., Qin, L., Swindells, M.B., Yamazaki, T., Ono, A.M., Kainosho, M., Inouye, M., and Ikura, M. (1998). NMR structure of the histidine kinase domain of the *E. coli* osmosensor EnvZ. Nature *396*, 88-92.

Tatsuno, I., Homma, M., Oosawa, K., and Kawagishi, I. (1996). Signaling by the *Escherichia coli* aspartate chemoreceptor Tar with a single cytoplasmic domain per dimer. Science *274*, 423-425.

Tatusova, T.A. and Madden, T.L. (1999). BLAST 2 sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol.Lett. 247-250.

Taylor, B.L. and Zhulin, I.B. (1999). PAS Domains: Internal Sensors of Oxygen, Redox Potential and Light. Micro.Mol.Biol Rev. (in press)

Terwilliger, T.C., Bogonez, E., Wang, E.A., and Koshland, D.E.J. (1983). Sites of methyl esterification on the aspartate receptor involved in bacterial chemotaxis. J.Biol.Chem. *258*, 9608-9611.

Terwilliger, T.C., Bollag, G.E., Sternberg, D.W.J., and Koshland, D.E.J. (1986). S-methyl glutathione synthesis is catalyzed by the cheR methyltransferase in *Escherichia coli*. J.Bacteriol. *165*, 958-963.

Terwilliger, T.C., Wang, J.Y., and Koshland, D.E.J. (1986). Kinetics of receptor modification. The multiply methylated aspartate receptors involved in bacterial chemotaxis. J.Biol.Chem. *261*, 10814-10820.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic.Acids.Res. *25*, 4876-4882.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic.Acids.Res. *22*, 4673-4680.

Thornton, J.M. (1998). The future of bioinformatics. In Trends Guide to Bioinformatics. S. Brenner and F. Lewitter, eds. (Oxford: Elsevier Science Ltd.), pp. 30-31.

Toker, A.S. and Macnab, R.M. (1997). Distinct regions of bacterial flagellar switch protein FliM interact with FliG, FliN and CheY. J.Mol.Biol. *273*, 623-634.

Volz, K. and Matsumura, P. (1991). Crystal structure of *Escherichia coli* CheY refined at 1.7-A resolution. J.Biol.Chem. *266*, 15511-15519.

Watson, J.D. and Crick, F.H.C. (1953). Molecular Structure of Nucleic Acids. Nature *171*, 737-739.

Welch, M., Chinardet, N., Mourey, L., Birck, C., and Samama, J.P. (1998). Structure of the CheY-binding domain of histidine kinase CheA in complex with CheY. Nat.Struct.Biol. *5*, 25-29.

Welch, M., Oosawa, K., Aizawa, S., and Eisenbach, M. (1993). Phosphorylation-dependent binding of a signal molecule to the flagellar switch of bacteria. Proc.Natl.Acad.Sci.U.S.A. *90*, 8787-8791.

West, A.H., Martinez-Hackert, E., and Stock, A.M. (1995). Crystal structure of the catalytic domain of the chemotaxis receptor methylesterase, CheB. J.Mol.Biol *250*, 276-290.

Whisstock, J.C. and Lesk, A.M. (1999). SH3 domains in prokaryotes. Trends Biochem.Sci. *24*, 132-133.

Wootton, J.C. and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. Methods Enzymol. *266:554-71*, 554-571.

Wu, J., Li, J., Li, G., Long, D.G., and Weis, R.M. (1996). The receptor binding site for the methyltransferase of bacterial chemotaxis is distinct from the sites of methylation. Biochemistry *35*, 4984-4993.

Yeh, J.I., Biemann, H.P., Pandit, J., Koshland, D.E., and Kim, S.H. (1993). The three-dimensional structure of the ligand-binding domain of a wild- type bacterial chemotaxis receptor. Structural comparison to the cross- linked mutant forms and conformational changes upon ligand binding. J.Biol.Chem. *268*, 9787-9792.

Yeh, J.I., Biemann, H.P., Prive, G.G., Pandit, J., Koshland, D.E.J., and Kim, S.H. (1996). High-resolution structures of the ligand binding domain of the wild- type bacterial aspartate receptor. J.Mol.Biol. *262*, 186-201.

Zhang, W., Brooun, A., McCandless, J., Banda, P., and Alam, M. (1996). Signal transduction in the archaeon *Halobacterium salinarium* is processed through three subfamilies of 13 soluble and membrane-bound transducer proteins. Proc.Natl.Acad.Sci.U.S.A. *93*, 4649-4654.

Zhang, W., Brooun, A., Mueller, M.M., and Alam, M. (1996). The primary structures of the Archaeon *Halobacterium salinarium* blue light receptor sensory rhodopsin II and its transducer, a methyl- accepting protein. Proc.Natl.Acad.Sci.U.S.A. *93*, 8230-8235.

Zhulin, I.B. (1999). Bacterial chemotaxis from the genomics point of view. Submitted

Zhulin, I.B. and Taylor, B.L. (1998). Correlation of PAS domains with electron transport-associated proteins in completely sequenced microbial genomes. Mol.Microbiol. *29*, 1522-1523.

Zhulin, I.B., Taylor, B.L., and Dixon, R. (1997). PAS domain S-boxes in Archea, Bacteria and sensors for oxygen and redox. Trends Biochem.Sci. *22*, 331-333.